

*An open text authored by the Actuarial Community*

---

# ***Loss Data Analytics***

## ***Second Edition***



---

# **Contents**

---

<b>Preface</b>	<b>9</b>
<b>1 Loss Data and Insurance Activities</b>	<b>1</b>
1.1 Data Driven Insurance Activities . . . . .	1
1.1.1 Nature and Relevance of Insurance . . . . .	1
1.1.2 Why Data Driven? . . . . .	2
1.1.3 Insurance Processes . . . . .	3
1.2 Insurance Company Operations . . . . .	6
1.2.1 Initiating Insurance . . . . .	7
1.2.2 Renewing Insurance . . . . .	8
1.2.3 Claims and Product Management . . . . .	9
1.2.4 Loss Reserving . . . . .	11
1.3 Case Study: Wisconsin Property Fund . . . . .	12
1.3.1 Fund Claims Variables: Frequency and Severity . . . . .	13
1.3.2 Fund Rating Variables . . . . .	14
1.3.3 Fund Operations . . . . .	19
1.4 Exercises . . . . .	22
1.5 Further Resources and Contributors . . . . .	27
<b>2 Introduction to Data Analytics</b>	<b>29</b>
2.1 Elements of Data Analytics . . . . .	29
2.1.1 Key Data Analytic Concepts . . . . .	30
2.1.2 Data versus Algorithmic Modeling . . . . .	32
2.2 Data Analysis Process . . . . .	33
2.3 Single Variable Analytics . . . . .	36
2.3.1 Variable Types . . . . .	36
2.3.2 Exploratory versus Confirmatory . . . . .	38
2.3.3 Model Construction . . . . .	39
2.3.4 Model Selection . . . . .	41
2.4 Analytics with Many Variables . . . . .	43
2.4.1 Supervised and Unsupervised Learning . . . . .	44
2.4.2 Algorithmic Modeling . . . . .	45
2.4.3 Data Modeling . . . . .	47
2.5 Data . . . . .	48

2.5.1	Data Types . . . . .	49
2.5.2	Data Structures and Storage . . . . .	49
2.5.3	Data Cleaning . . . . .	50
2.5.4	Big Data Analysis . . . . .	51
2.5.5	Ethical Issues . . . . .	52
2.6	Further Resources and Contributors . . . . .	52
2.6.1	Technical Supplement: Multivariate Exploratory Analysis . . . . .	53
2.6.2	Tree-based Models . . . . .	54
2.6.3	Technical Supplement: Some R Functions . . . . .	55
<b>3</b>	<b>Frequency Modeling</b>	<b>57</b>
3.1	Frequency Distributions . . . . .	58
3.1.1	How Frequency Augments Severity Information . . . . .	58
3.2	Basic Frequency Distributions . . . . .	60
3.2.1	Foundations . . . . .	60
3.2.2	Moment and Probability Generating Functions . . . . .	62
3.2.3	Important Frequency Distributions . . . . .	64
3.3	The $(a, b, 0)$ Class . . . . .	68
3.4	Estimating Frequency Distributions . . . . .	71
3.4.1	Parameter Estimation . . . . .	71
3.4.2	Frequency Distributions MLE . . . . .	74
3.5	Other Frequency Distributions . . . . .	82
3.5.1	Zero Truncation or Modification . . . . .	83
3.6	Mixture Distributions . . . . .	85
3.7	Real Data Example . . . . .	89
3.8	Exercises . . . . .	90
3.9	Further Resources and Contributors . . . . .	94
3.9.1	TS 3.A. R Code for Plots . . . . .	95
<b>4</b>	<b>Modeling Loss Severity</b>	<b>99</b>
4.1	Basic Distributional Quantities . . . . .	99
4.1.1	Moments and Moment Generating Functions . . . . .	100
4.1.2	Quantiles . . . . .	103
4.2	Continuous Distributions for Modeling Loss Severity . . . . .	104
4.2.1	Gamma Distribution . . . . .	104
4.2.2	Pareto Distribution . . . . .	106
4.2.3	Weibull Distribution . . . . .	108
4.2.4	The Generalized Beta Distribution of the Second Kind	111
4.3	Methods of Creating New Distributions . . . . .	112
4.3.1	Functions of Random Variables and their Distributions	112
4.3.2	Mixture Distributions for Severity . . . . .	117

<i>Contents</i>	5
4.4 Estimating Loss Distributions . . . . .	121
4.4.1 Nonparametric Estimation . . . . .	121
4.4.2 Parametric Estimation . . . . .	130
4.5 Exercises with a Practical Focus . . . . .	142
4.6 Further Resources and Contributors . . . . .	143
<b>5 Modeling Claim Severity</b>	<b>145</b>
5.1 Coverage Modifications . . . . .	145
5.1.1 Policy Deductibles . . . . .	146
5.1.2 Policy Limits . . . . .	150
5.1.3 Coinsurance and Inflation . . . . .	152
5.1.4 Reinsurance . . . . .	154
5.2 Parametric Estimation using Modified Data . . . . .	156
5.2.1 Parametric Estimation using Grouped Data . . . . .	157
5.2.2 Censored Data . . . . .	158
5.2.3 Truncated Data . . . . .	159
5.2.4 Parametric Estimation using Censored and Truncated Data . . . . .	162
5.3 Nonparametric Estimation using Modified Data . . . . .	165
5.3.1 Grouped Data . . . . .	165
5.3.2 Plug-in Principle . . . . .	167
5.3.3 Right-Censored Empirical Distribution Function . . . . .	169
5.4 Further Resources and Contributors . . . . .	174
<b>6 Model Selection</b>	<b>175</b>
6.1 Tools for Model Selection and Diagnostics . . . . .	175
6.1.1 Graphical Comparison of Distributions . . . . .	176
6.1.2 Statistical Comparison of Distributions . . . . .	180
6.2 Iterative Model Selection . . . . .	184
6.3 Model Selection Based on a Training Dataset . . . . .	185
6.4 Model Selection Based on a Test Dataset . . . . .	186
6.5 Model Selection Based on Cross-Validation . . . . .	192
6.6 Model Selection for Modified Data . . . . .	196
6.7 Further Resources and Contributors . . . . .	198
<b>7 Aggregate Loss Models</b>	<b>199</b>
7.1 Introduction . . . . .	199
7.2 Individual Risk Model . . . . .	201
7.2.1 Moments and Distribution . . . . .	201
7.2.2 Aggregate Loss Distribution . . . . .	205
7.3 Collective Risk Model . . . . .	209
7.3.1 Moments and Distribution . . . . .	209

7.3.2	Stop-loss Insurance . . . . .	217
7.3.3	Closed-form Distributions . . . . .	220
7.3.4	Tweedie Distribution . . . . .	222
7.4	Computing the Aggregate Claims Distribution . . . . .	224
7.4.1	Recursive Method . . . . .	224
7.4.2	Simulation . . . . .	227
7.5	Effects of Coverage Modifications . . . . .	229
7.5.1	Impact of Exposure on Frequency . . . . .	230
7.5.2	Impact of Deductibles on Claim Frequency . . . . .	231
7.5.3	Impact of Policy Modifications on Aggregate Claims .	236
7.6	Further Resources and Contributors . . . . .	240
<b>8</b>	<b>Simulation and Resampling</b>	<b>245</b>
8.1	Random Number Generation . . . . .	245
8.1.1	Generating Independent Uniform Observations . . . .	246
8.1.2	Inverse Transform Method . . . . .	248
8.1.3	Ready-made Random Number Generators . . . . .	252
8.1.4	Simulating from Complex Distributions . . . . .	253
8.1.5	Importance Sampling . . . . .	254
8.2	Computing Distribution Parameters . . . . .	256
8.2.1	Simulating Parameters . . . . .	257
8.2.2	Determining the Number of Simulations . . . . .	258
8.2.3	Simulation and Statistical Inference . . . . .	261
8.3	Bootstrapping and Resampling . . . . .	265
8.3.1	Bootstrap Foundations . . . . .	265
8.3.2	Bootstrap Precision: Bias, Standard Deviation, and Mean Square Error . . . . .	267
8.3.3	Confidence Intervals . . . . .	272
8.3.4	Parametric Bootstrap . . . . .	274
8.4	Model Selection and Cross-Validation . . . . .	276
8.4.1	k-Fold Cross-Validation . . . . .	277
8.4.2	Leave-One-Out Cross-Validation . . . . .	278
8.4.3	Cross-Validation and Bootstrap . . . . .	280
8.5	Further Resources and Contributors . . . . .	281
<b>9</b>	<b>Bayesian Statistics and Modeling</b>	<b>283</b>
9.1	A Gentle Introduction to Bayesian Statistics . . . . .	284
9.1.1	Bayesian versus Frequentist Statistics . . . . .	284
9.1.2	A Brief History Lesson . . . . .	289
9.1.3	Bayes' Rule . . . . .	290
9.1.4	An Introductory Example of Bayes' Rule . . . . .	294
9.2	Building Blocks of Bayesian Statistics . . . . .	295

9.2.1	Posterior Distribution . . . . .	297
9.2.2	Likelihood Function . . . . .	300
9.2.3	Prior Distribution . . . . .	301
9.3	Conjugate Families . . . . .	308
9.3.1	The Beta–Binomial Conjugate Family . . . . .	309
9.3.2	The Gamma–Poisson Conjugate Family . . . . .	314
9.3.3	The Normal–Normal Conjugate Family . . . . .	317
9.3.4	Criticism of Conjugate Family Models . . . . .	321
9.4	Posterior Simulation . . . . .	321
9.4.1	Introduction to Markov Chain Monte Carlo Methods . . . . .	321
9.4.2	The Gibbs Sampler . . . . .	322
9.4.3	The Metropolis–Hastings Algorithm . . . . .	331
9.4.4	Markov Chain Diagnostics . . . . .	336
9.5	Bayesian Statistics in Practice . . . . .	340
9.6	Further Resources and Contributors . . . . .	342
<b>10</b>	<b>Premium Foundations</b>	<b>345</b>
10.1	Introduction to Ratemaking . . . . .	345
10.2	Data Sources . . . . .	347
10.3	Claims . . . . .	348
10.3.1	Estimated Ultimate Claims . . . . .	349
10.3.2	Adjustments to Claims and Allocated Claims Adjustment Expenses . . . . .	350
10.4	Exposures . . . . .	351
10.4.1	Criteria for Choosing an Exposure . . . . .	352
10.4.2	Written and Earned Exposures . . . . .	353
10.4.3	Adjustments to Exposures . . . . .	353
10.5	Pure Premiums . . . . .	353
10.5.1	Experience Period . . . . .	354
10.5.2	Expected Pure Premium . . . . .	354
10.6	Non-Claim Expenses . . . . .	355
10.7	Investment Income . . . . .	355
10.7.1	Investment Income on Policyholder Cash Flows . . . . .	356
10.7.2	Investment Income on Surplus . . . . .	357
10.7.3	The Underwriting Profit Provisions . . . . .	357
10.8	The Premium Equation . . . . .	358
10.9	Pricing Principles . . . . .	358
10.9.1	Premium Principles . . . . .	359
10.9.2	Properties of Premium Principles . . . . .	360
10.10	Reviewing Rate Adequacy . . . . .	361
10.10.1	The Loss Ratio Method . . . . .	361
10.10.2	Target Loss Ratio . . . . .	362

10.10.3 Experience Period Loss Ratios . . . . .	362
10.10.4 Adjustments to Loss . . . . .	362
10.10.5 Premium On-Level Adjustment . . . . .	363
10.10.6 Premium Trend . . . . .	364
10.10.7 Credibility . . . . .	364
10.11 Further Resources and Contributors . . . . .	365

---

# Preface

---

Date: 14 September 2024

## Book Description

**Loss Data Analytics** is an interactive, online, freely available text.

- The online version contains many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote *deeper learning*.
- A subset of the book is available for *offline reading* in pdf and EPUB formats.
- The online text will be available in multiple languages to promote access to a *worldwide audience*.

## What will success look like?

The online text will be freely available to a worldwide audience. The online version will contain many interactive objects (quizzes, computer demonstrations, interactive graphs, video, and the like) to promote deeper learning. Moreover, a subset of the book will be available in pdf format for low-cost printing. The online text will be available in multiple languages to promote access to a worldwide audience.

## How will the text be used?

This book will be useful in actuarial curricula worldwide. It will cover the loss data learning objectives of the major actuarial organizations. Thus, it will be suitable for classroom use at universities as well as for use by independent learners seeking to pass professional actuarial examinations. Moreover, the text will also be useful for the continuing professional development of actuaries and other professionals in insurance and related financial risk management industries.

## Why is this good for the profession?

An online text is a type of open educational resource (OER). One important benefit of an OER is that it equalizes access to knowledge, thus permitting a broader community to learn about the actuarial profession. Moreover, it

has the capacity to engage viewers through active learning that deepens the learning process, producing analysts more capable of solid actuarial work.

Why is this good for students and teachers and others involved in the learning process? Cost is often cited as an important factor for students and teachers in textbook selection (see a recent post on the [\\$400 textbook](#)). Students will also appreciate the ability to “carry the book around” on their mobile devices.

### Why loss data analytics?

The intent is that this type of resource will eventually permeate throughout the actuarial curriculum. Given the dramatic changes in the way that actuaries treat data, loss data seems like a natural place to start. The idea behind the name *loss data analytics* is to integrate classical loss data models from applied probability with modern analytic tools. In particular, we recognize that big data (including social media and usage based insurance) are here to stay and that high speed computation is readily available.

### Project Goal

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. To get involved, please visit our [Open Actuarial Textbooks Project Site](#).

---

### Acknowledgements

Edward Frees acknowledges the John and Anne Oros Distinguished Chair for Inspired Learning in Business which provided seed money to support the project. Frees and his Wisconsin colleagues also acknowledge a Society of Actuaries Center of Excellence Grant that provided funding to support work in dependence modeling and health initiatives. Wisconsin also provided an education innovation grant that provided partial support for the many students who have worked on this project.

We acknowledge the Society of Actuaries for permission to use problems from their examinations.

We thank Rob Hyndman, Monash University, for allowing us to use his excellent style files to produce the online version of the book.

We thank Yihui Xie and his colleagues at [Rstudio](#) for the [R bookdown](#) package that allows us to produce this book.

We also wish to acknowledge the support and sponsorship of the [International](#)

Association of Black Actuaries in our joint efforts to provide actuarial educational content to all.



---

## Contributors

The project goal is to have the actuarial community author our textbooks in a collaborative fashion. The following contributors have taken a leadership role in developing *Loss Data Analytics*.



**Zeinab Amin**

- **Zeinab Amin** is a Professor at the Department of Mathematics and Actuarial Science and Associate Provost for Assessment and Accreditation at the American University in Cairo (AUC). Amin holds a PhD in Statistics and is an Associate of the Society of Actuaries. Amin is the recipient of the 2016 Excellence in Academic Service Award and the 2009 Excellence in Teaching Award from AUC. Amin has designed and taught a variety of statistics and actuarial science courses. Amin's current area of research includes quantitative risk assessment, reliability assessment, general statistical modelling, and Bayesian statistics.
  - **Katrien Antonio**, KU Leuven
-



**Jean-François Bégin**

- **Jean-François Bégin** is an Assistant Professor in the Department of Statistics and Actuarial Science at Simon Fraser University in British Columbia, Canada. Bégin holds a PhD in Financial Engineering from HEC Montréal, Canada, and is a Fellow of the Society of Actuaries and of the Canadian Institute of Actuaries. His current research interests include financial modelling, financial econometrics, Bayesian statistics, filtering methods, credit risk, option pricing, and pension economics. Bégin has designed and taught a variety of actuarial finance and actuarial communication courses.
- **Jan Beirlant**, KU Leuven



**Arthur Charpentier**

- **Arthur Charpentier** is a professor in the Department of Mathematics at the Université du Québec à Montréal. Prior to that, he worked at a large general insurance company in Hong Kong, China, and the French Federation of Insurers in Paris, France. He received a MS on mathematical economics at Université Paris Dauphine and a MS in actuarial science at ENSAE (National School of Statistics) in Paris, and a PhD degree from KU Leuven, Belgium. His research interests include econometrics, applied probability and actuarial science. He has published several books (the most recent one on *Computational Actuarial Science with R*, CRC) and papers on a variety of topics. He is a Fellow of the French Institute of Actuaries,

and was in charge of the ‘Data Science for Actuaries’ program from 2015 to 2018.



**Curtis Gary Dean**

- **Curtis Gary Dean** is the Lincoln Financial Distinguished Professor of Actuarial Science at Ball State University. He is a Fellow of the Casualty Actuarial Society and a CFA charterholder. He has extensive practical experience as an actuary at American States Insurance, SAFECO, and Travelers. He has served the CAS and actuarial profession as chair of the Examination Committee, first editor-in-chief for *Variance: Advancing the Science of Risk*, and as a member of the Board of Directors and the Executive Council. He contributed a chapter to *Predictive Modeling Applications in Actuarial Science* published by Cambridge University Press.



**Edward (Jed) Frees**

- **Edward (Jed) Frees** is an emeritus professor, formerly the Hickman-Larson Chair of Actuarial Science at the University of Wisconsin-Madison. He is a Fellow of both the Society of Actuaries and the American Statistical Association. He has published extensively (a four-time winner of the Halmstad and Prize for best paper published in the actuarial literature) and has written three books. He also is a co-editor of the two-volume series *Predictive Modeling Applications in Actuarial Science* published by Cambridge University Press.

**Guojun Gan**

- **Guojun Gan** is an associate professor in the Department of Mathematics at the University of Connecticut, where he has been since August 2014. Prior to that, he worked at a large life insurance company in Toronto, Canada for six years. He received a BS degree from Jilin University, Changchun, China, in 2001 and MS and PhD degrees from York University, Toronto, Canada, in 2003 and 2007, respectively. His research interests include data mining and actuarial science. He has published several books and papers on a variety of topics, including data clustering, variable annuity, mathematical finance, applied statistics, and VBA programming.

**Lisa Gao**

- **Lisa Gao** is a PhD candidate in the Risk and Insurance department at the University of Wisconsin-Madison. She holds a BMath in Actuarial Science and Statistics from the University of Waterloo and is an Associate of the Society of Actuaries.
- **José Garrido**, Concordia University

**Lei (Larry) Hua**

- **Lei (Larry) Hua** is an Associate Professor of Actuarial Science at Northern Illinois University. He earned a PhD degree in Statistics from the University of British Columbia. He is an Associate of the Society of Actuaries. His research work focuses on multivariate dependence modeling for non-Gaussian phenomena and innovative applications for financial and insurance industries.



**Noriszura Ismail**

- **Noriszura Ismail** is a Professor and Head of Actuarial Science Program, Universiti Kebangsaan Malaysia (UKM). She specializes in Risk Modelling and Applied Statistics. She obtained her BSc and MSc (Actuarial Science) in 1991 and 1993 from University of Iowa, and her PhD (Statistics) in 2007 from UKM. She also passed several papers from Society of Actuaries in 1994. She has received several research grants from Ministry of Higher Education Malaysia (MOHE) and UKM, totaling about MYR1.8 million. She has successfully supervised and co-supervised several PhD students (13 completed and 11 on-going). She currently has about 180 publications, consisting of 88 journals and 95 proceedings.



**Joseph H.T. Kim**

- **Joseph H.T. Kim**, Ph.D., FSA, CERA, is Associate Professor of Applied Statistics at Yonsei University, Seoul, Korea. He holds a Ph.D. degree in Actuarial Science from the University of Waterloo, at which he taught as Assistant Professor. He also worked in the life insurance industry. He has published papers in *Insurance Mathematics and Economics*, *Journal of Risk and Insurance*, *Journal of Banking and Finance*, *ASTIN Bulletin*, and *North American Actuarial Journal*, among others.



### Nii-Armah Okine

- **Nii-Armah Okine** is an assistant professor at the Mathematical Sciences Department at Appalachian State University. He holds a Ph.D. in Business (Actuarial Science) from the University of Wisconsin - Madison and obtained his master's degree in Actuarial science from Illinois State University. His research interest includes micro-level reserving, joint longitudinal-survival modeling, dependence modeling, micro-insurance, and machine learning.

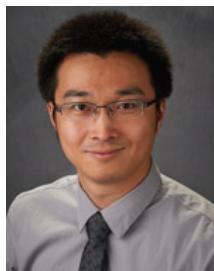


### Rajesh (Raj) Sahasrabuddhe

- **Rajesh (Raj) Sahasrabuddhe** is a Partner and Philadelphia Office Leader with Oliver Wyman Actuarial Consulting. Raj is a Fellow of the Casualty Actuarial Society (CAS), an Associate of the Canadian Institute of Actuaries, and a Member of the American Academy of Actuaries. Raj has been an active volunteer with CAS Admissions committees throughout his career, including a term as Chairperson of the Syllabus Committee from 2010 to 2013. He currently serves on the MAS-II Examination Committee. He has authored or co-authored papers that have appeared on syllabi for both the CAS and Society of Actuaries.

**Emine Selin Sarıdaş**

- **Emine Selin Sarıdaş** is a doctoral candidate in the Statistics department of Mimar Sinan University. She holds a bachelor degree in Actuarial Science with a minor in Economics and a master degree in Actuarial Science from Hacettepe University. Her research interest includes dependence modeling, regression, loss models and life contingencies.

**Peng Shi**

- **Peng Shi** is an associate professor in the Risk and Insurance Department at the Wisconsin School of Business. He is also the Charles & Laura Albright Professor in Business and Finance. Professor Shi is an Associate of the Casualty Actuarial Society (ACAS) and a Fellow of the Society of Actuaries (FSA). He received a Ph.D. in actuarial science from the University of Wisconsin-Madison. His research interests are problems at the intersection of insurance and statistics. He has won several research awards, including the Charles A. Hachemeister Prize, the Ronald Bornhuetter Loss Reserve Prize, and the American Risk and Insurance Association Prize.



**Nariankadu D. Shyamalkumar (Shyamal)**

- **Nariankadu D. Shyamalkumar (Shyamal)** is an associate professor in the Department of Statistics and Actuarial Science at The University of Iowa. He is an Associate of the Society of Actuaries, and has volunteered in various elected and non-elected roles within the SoA. Having a broad theoretical interest as well as interest in computing, he has published in prominent actuarial, computer science, probability theory, and statistical journals. Moreover, he has worked in the financial industry, and since then served as an independent consultant to the insurance industry. He has experience educating actuaries in both Mexico and the US, serving in the roles of directing an undergraduate program, and as a graduate adviser for both masters and doctoral students.



**Jianxi Su**

- **Jianxi Su** is an Assistant Professor at the Department of Statistics at Purdue University. He is the Associate Director of Purdue's Actuarial Science. Prior to joining Purdue in 2016, he completed the PhD at York University (2012-2015). He obtained the Fellow of the Society of Actuaries (FSA) in 2017. His research expertise are in dependence modelling, risk management, and pricing. During the PhD candidature, Jianxi also worked as a research associate at the Model Validation and ORSA Implementation team of Sun Life Financial (Toronto office).

**Chong It Tan**

- **Chong It Tan** is a senior lecturer at Macquarie University in Australia, where he has served as the undergraduate actuarial program director since 2018. He obtained his PhD in 2015 from Nanyang Technological University in Singapore. He is a fully qualified actuary, holding the credentials from both the US Society of Actuaries and Australian Actuaries Institute. His major research interests are mortality modelling, longevity risk management and bonus-malus systems.

**Tim Verdonck**

- **Tim Verdonck** is associate professor at the University of Antwerp. He has a degree in Mathematics and a PhD in Science: Mathematics, obtained at the University of Antwerp. During his PhD he successfully took the Master in Insurance and the Master in Financial and Actuarial Engineering, both at KU Leuven. His research focuses on the adaptation and application of robust statistical methods for insurance and finance data.



### Krupa Viswanathan

- **Krupa Viswanathan** is an Associate Professor in the Risk, Insurance and Healthcare Management Department in the Fox School of Business, Temple University. She is an Associate of the Society of Actuaries. She teaches courses in Actuarial Science and Risk Management at the undergraduate and graduate levels. Her research interests include corporate governance of insurance companies, capital management, and sentiment analysis. She received her Ph.D. from The Wharton School of the University of Pennsylvania.
- 

### Reviewers

Our goal is to have the actuarial community author our textbooks in a collaborative fashion. Part of the writing process involves many reviewers who generously donated their time to help make this book better. They are:

- Yair Babab
- David Back, Liberty Mutual
- Chunsheng Ban, Ohio State University
- Vytautas Brazauskas, University of Wisconsin - Milwaukee
- Yvonne Chueh, Central Washington University
- Chun Yong Chew, Universiti Tunku Abdul Rahman (UTAR)
- Benjamin Côté, Université Laval
- Eren Dodd, University of Southampton
- Gordon Enderle, University of Wisconsin - Madison
- Rob Erhardt, Wake Forest University
- Runhun Feng, University of Illinois
- Brian Hartman, Brigham Young University
- Liang (Jason) Hong, University of Texas at Dallas
- Fei Huang, Australian National University
- Hirokazu (Iwahiro) Iwasawa

- Himchan Jeong, University of Connecticut
- Min Ji, Towson University
- Paul Herbert Johnson, University of Wisconsin - Madison
- Dalia Khalil, Cairo University
- Samuel Kolins, Lebonan Valley College
- Andrew Kwon-Nakamura, Zurich North America
- Ambrose Lo, University of Iowa
- Mélina Mailhot, Concordia University
- Mark Maxwell, University of Texas at Austin
- Tatjana Miljkovic, Miami University
- Bell Ouelega, American University in Cairo
- Zhiyu (Frank) Quan, University of Connecticut
- Jiandong Ren, Western University
- Margie Rosenberg, University of Wisconsin - Madison
- Rajesh V. Sahasrabuddhe, Oliver Wyman
- Sherly Paola Alfonso Sanchez, Universidad Nacional de Colombia
- Ranee Thiagarajah, Illinois State University
- Ping Wang, Saint Johns University
- Chengguo Weng, University of Waterloo
- Toby White, Drake University
- Michelle Xia, Northern Illinois University
- Di (Cindy) Xu, University of Nebraska - Lincoln
- Lina Xu, Columbia University
- Lu Yang, University of Amsterdam
- Chun Yong
- Jorge Yslas, University of Copenhagen
- Jeffrey Zheng, Temple University
- Hongjuan Zhou, Arizona State University

### **Other Collaborators**

- Alyaa Nuval Binti Othman, Aisha Nuval Binti Othman, and Khairina (Rina) Binti Ibrahim were three of many students at the Univeristy of Wiscinson-Madison that helped with the text over the years.
- Maggie Lee, Macquarie University, and Anh Vu (then at University of New South Wales) contributed the end of the section quizzes.
- Jeffrey Zheng, Temple University, Lu Yang (University of Amsterdam), and Paul Johnson, University of Wisconsin-Madison, led the work on the glossary.

---

## Version Number

- This is **Version 2.0**, October 2024. Edited by Hélène Cossette, Edward (Jed) Frees, Brian Hartman, and Tim Higgins.
- Version 1.1, August 2020. Edited by Edward (Jed) Frees and Paul Johnson.
- Version 1.0, January 2020, was edited by Edward (Jed) Frees.

You can also access pdf and epub (current and older) versions of the text in our [Offline versions of the text](#).

---

## For our Readers

We hope that you find this book worthwhile and even enjoyable. For your convenience, at our [Github Landing site \(<https://openacttexts.github.io/>\)](#), you will find links to the book that you can (freely) download for offline reading, including a pdf version (for Adobe Acrobat) and an EPUB version suitable for mobile devices. [Data](#) for running our examples are available at the same site.

In developing this book, we are emphasizing the [online version](#) that has lots of great features such as a glossary, code and solutions to examples that you can be revealed interactively. For example, you will find that the statistical code is hidden and can only be seen by clicking on terms such as

We hide the code because we don't want to insist that you use the R statistical software (although we like it). Still, we encourage you to try some statistical code as you read the book – we have opted to make it easy to learn R as you go. We have set up a separate [R Code for Loss Data Analytics](#) site to explain more of the details of the code.

Like any book, we have a set of notations and conventions. It will probably save you time if you regularly visit our Appendix Chapter ?? to get used to ours.

Freely available, interactive textbooks represent a new venture in actuarial education and we need your input. Although a lot of effort has gone into the development, we expect hiccoughs. Please let your instructor know about opportunities for improvement, write us through our project site, or contact chapter contributors directly with suggested improvements.

---

This work is licensed under a Creative Commons Attribution 4.0 International License.



# 1

---

## *Loss Data and Insurance Activities*

---

*Chapter Preview.* This book introduces readers to methods of analyzing insurance data. Section 1.1 begins with a discussion of why the use of data is important in the insurance industry. Section 1.2 gives a general overview of the purposes of analyzing insurance data which is reinforced in the Section 1.3 case study. Naturally, there is a huge gap between the broad goals summarized in the overview and a case study application; this gap is covered through the methods and techniques of data analysis covered in the rest of the text.

---

### **1.1 Data Driven Insurance Activities**

---

In this section, you learn how to:

- Summarize the importance of insurance to consumers and the economy
  - Describe the role that data plays in managing insurance activities
  - Identify data generating events associated with the timeline of a typical insurance contract
- 

#### **1.1.1 Nature and Relevance of Insurance**

This book introduces the process of using data to make decisions in an insurance context. It does not assume that readers are familiar with insurance but introduces insurance concepts as needed. Insurance is the exchange of a certain amount, known as a premium, for a promise to compensate another party upon the occurrence of an insured event.

If you are new to insurance, then it is probably easiest to think about an insurance policy that covers the contents of an apartment or house that you are renting (known as renters insurance) or the contents and property of a building that is owned by you or a friend (known as homeowners insurance). Another common example is automobile insurance. In the event of an accident,

this policy may cover damage to your vehicle, damage to other vehicles in the accident, as well as medical expenses of those injured in the accident.

One way to think about the nature of insurance is who buys it. Renters, homeowners, and auto insurance are examples of personal insurance in that these are policies issued to people. Businesses also buy insurance, such as coverage on their properties, and this is known as commercial insurance. The seller, an insurance company, is also known as an insurer. Even insurance companies need insurance; this is known as reinsurance.

Another way to think about the nature of insurance is the type of risk being covered. In the U.S., policies such as renters and homeowners are known as property insurance whereas a policy such as auto that covers medical damages to people is known as casualty insurance. In the rest of the world, these are both known as non-life or general insurance, to distinguish them from life insurance.

Both life and non-life insurances are important components of the world economy. The [The Organization for Economic Cooperation and Development \(OECD\)](#) estimates that direct insurance premiums in the OECD (Organization for Economic Cooperation and Development) countries for 2020 was 2,520,220 for life and 2,704,799 for non-life; these figures are in *millions of U.S. dollars*. The total represents 9.447% of the OECD gross domestic product (GDP). As examples, premiums accounted for 30.9% of GDP in Luxembourg and 17.0% of GDP in Chinese Taipei (the two highest in the study) and represented 12.5% of GDP in the United States. Both life and non-life insurances represent important economic activities.

Insurance affects the financial livelihoods of many and, by almost any measure, insurance is a major economic activity. As noted earlier, on a global level insurance premiums comprised nearly 9.5% of GDP in 2020. On a personal level, almost everyone owning a home has insurance to protect themselves in the event of a fire, hailstorm, or some other calamitous event. Almost every country requires insurance for those driving a car. In sum, insurance plays an important role in the economies of nations and the lives of individuals.

### 1.1.2 Why Data Driven?

Insurance is a data-driven industry. Like all major corporations and organizations, insurers use data when trying to decide how much to pay employees, how many employees to retain, how to market their services and products, how to forecast financial trends, and so on. These represent general areas of activities that are not specific to the insurance industry. Although each industry has its own data nuances and needs, the collection, analysis and use of

data is an activity shared by all, from the internet giants to a small business, by public and governmental organizations, and is not specific to the insurance industry. You will find that the data collection and analysis methods and tools introduced in this text are relevant for all.

In any data-driven industry, deriving and extracting information from data is critical. Making data-driven business decisions has been described as business analytics, business intelligence, and data science. These terms, among others, are sometimes used interchangeably and sometimes refer to distinct applications. *Business intelligence* may focus on processes of collecting data, often through databases and data warehouses, whereas *business analytics* utilizes tools and methods for statistical analyses of data. In contrast to these two terms that emphasize business applications, the term *data science* can encompass broader data related applications in many scientific domains. For our purposes, we use the term analytics to refer to the process of using data to make decisions. This process involves gathering data, understanding concepts and models of uncertainty, making general inferences, and communicating results. Chapter 2 describes data analytics in further detail.

When introducing methods in this text, we focus on **loss data** that arise from, or are related to, obligations in insurance contracts. This could be the amount of damage to one's apartment under a renter's insurance agreement, the amount needed to compensate someone that you hurt in a driving accident, and the like. We call such payments an insurance claim. With this focus, we are able to introduce and directly use generally applicable statistical tools and techniques.

### 1.1.3 Insurance Processes

Yet another way to think about the nature of insurance is by the duration of an insurance contract, known as the term. This text will focus on short-term insurance contracts. By short-term, we mean contracts where the insurance coverage is typically provided for a year or six months. Most non-life commercial and personal contracts are for a year so that is our default duration. An important exception is U.S. auto policies that are often six months in length.

In contrast, we typically think of life insurance as a long-term contract where the default is to have a multi-year contract. For example, if a person 25 years old purchases a whole life policy that pays upon death of the insured and that person does not die until age 100, then the contract is in force for 75 years.

There are other important differences between life and non-life products. In life insurance, the benefit amount is often stipulated in the contract provisions. In contrast, most non-life contracts provide for compensation of insured losses

which are unknown before the accident. (There are usually limits placed on the compensation amounts.) In a life insurance contract that stretches over many years, the time value of money plays a prominent role. In a non-life contract, the random amount of compensation takes priority.

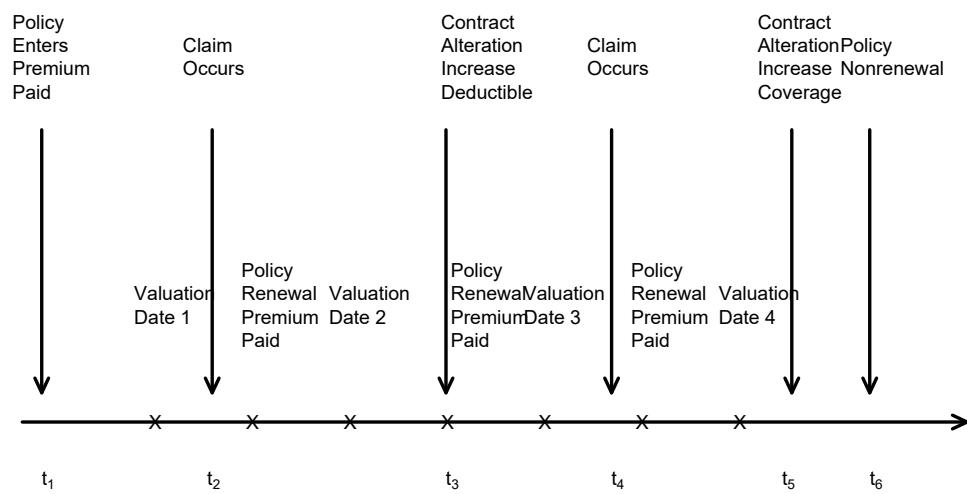
In both life and non-life insurances, the frequency of claims is very important. For many life insurance contracts, the insured event (such as death) happens only once. In contrast, for non-life insurances such as automobile, it is common for individuals (especially young male drivers) to get into more than one accident during a year. So, our models need to reflect this observation; we introduce different frequency models that you may also see when studying life insurance.

For short-term insurance, the framework of the probabilistic model is straightforward. We think of a one-period model (the period length, e.g., one year, will be specified in the situation).

- At the beginning of the period, the insured pays the insurer a known premium that is agreed upon by both parties to the contract.
- At the end of the period, the insurer reimburses the insured for a (possibly multivariate) random loss.

This framework will be developed as we proceed; but we first focus on integrating this framework with concerns about how the data may arise. From an insurer's viewpoint, contracts may be only for a year but they tend to be renewed. Moreover, payments arising from claims during the year may extend well beyond a single year. One way to describe the data arising from operations of an insurance company is to use a timeline granular approach. A **process** approach provides an overall view of the events occurring during the life of an insurance contract, and their nature – random or planned, loss events (claims) and contract changes events, and so forth. In this micro oriented view, we can think about what happens to a contract at various stages of its existence.

Figure 1.1 traces a timeline of a typical insurance contract. Throughout the life of the contract, the company regularly processes events such as premium collection and valuation, described in Section 1.2; these are marked with an **x** on the timeline. Non-regular and unanticipated events also occur. To illustrate,  $t_2$  and  $t_4$  mark the event of an insurance claim (some contracts, such as life insurance, can have only a single claim). Times  $t_3$  and  $t_5$  mark events when a policyholder wishes to alter certain contract features, such as the choice of a deductible or the amount of coverage. From a company perspective, one can even think about the contract initiation (arrival, time  $t_1$ ) and contract termination (departure, time  $t_6$ ) as uncertain events. (Alternatively, for some purposes, you may condition on these events and treat them as certain.)



**FIGURE 1.1: Timeline of a Typical Insurance Policy.** Arrows mark the occurrences of random events. Each x marks the time of scheduled events that are typically non-random.

## 1.2 Insurance Company Operations

---

In this section, you learn how to:

- Describe five major operational areas of insurance companies.
  - Identify the role of data and analytics opportunities within each operational area.
- 

Armed with insurance data, the end goal is to use data to make decisions. We will learn more about methods of analyzing and extrapolating data in future chapters. To begin, let us think about why we want to do the analysis. We take the insurance company's viewpoint (not the insured person) and introduce ways of bringing money in, paying it out, managing costs, and making sure that we have enough money to meet obligations. The emphasis is on insurance-specific operations rather than on general business activities such as advertising, marketing, and human resources management.

Specifically, in many insurance companies, it is customary to aggregate detailed insurance processes into larger operational units; many companies use these functional areas to segregate employee activities and areas of responsibilities. Actuaries, other financial analysts, and insurance regulators work within these units and use data for the following activities:

1. **Initiating Insurance.** At this stage, the company makes a decision as to whether or not to take on a risk (the underwriting stage) and assign an appropriate premium (or rate). Insurance analytics has its actuarial roots in *ratemaking*, where analysts seek to determine the right price for the right risk.
2. **Renewing Insurance.** Many contracts, particularly in general insurance, have relatively short durations such as 6 months or a year. Although there is an implicit expectation that such contracts will be renewed, the insurer has the opportunity to decline coverage and to adjust the premium. Analytics is also used at this policy renewal stage where the goal is to retain profitable customers.
3. **Claims Management.** Analytics has long been used in (1) detecting and preventing claims fraud, (2) managing claim costs, including identifying the appropriate support for claims handling expenses, as well as (3) understanding excess layers for reinsurance and retention.
4. **Loss Reserving.** Analytic tools are used to provide management

with an appropriate estimate of future obligations and to quantify the uncertainty of those estimates.

5. **Solvency and Capital Allocation.** Deciding on the requisite amount of capital and on ways of allocating capital among alternative investments are also important analytics activities. Companies must understand how much capital is needed so that they have sufficient flow of cash available to meet their obligations at the times they are expected to materialize (solvency). This is an important question that concerns not only company managers but also customers, company shareholders, regulatory authorities, as well as the public at large. Related to issues of how much capital is the question of how to allocate capital to differing financial projects, typically to maximize an investor's return. Although this question can arise at several levels, insurance companies are typically concerned with how to allocate capital to different lines of business within a firm and to different subsidiaries of a parent firm.

Although data represent a critical component of solvency and capital allocation, other components including the local and global economic framework, the financial investments environment, and quite specific requirements according to the regulatory environment of the day, are also important. Because of the background needed to address these components, we do not address solvency, capital allocation, and regulation issues in this text.

Nonetheless, for all operating functions, we emphasize that analytics in the insurance industry is not an exercise that a small group of analysts can do by themselves. It requires an insurer to make significant investments in their information technology, marketing, underwriting, and actuarial functions. As these areas represent the primary end goals of the analysis of data, additional background on each operational unit is provided in the following subsections.

### 1.2.1 Initiating Insurance

Setting the price of an insurance product can be a perplexing problem. This is in contrast to other industries such as manufacturing where the cost of a product is (relatively) known and provides a benchmark for assessing a market demand price. Similarly, in other areas of financial services, market prices are available and provide the basis for a market-consistent pricing structure of products. However, for many lines of insurance, the cost of a product is uncertain and market prices are unavailable. Expectations of the random cost is a reasonable place to start for a price. (If you have studied finance, then you will recall that an expectation is the optimal price for a risk-neutral insurer.) It has been traditional in insurance pricing to begin with the expected cost.

Insurers then add margins to this, to account for the product's riskiness, expenses incurred in servicing the product, and an allowance for profit/surplus of the company.

Use of expected costs as a foundation for pricing is prevalent in some lines of the insurance business. These include automobile and homeowners insurance. For these lines, analytics has served to sharpen the market by making the calculation of the product's expected cost more precise. The increasing availability of the internet to consumers has also promoted transparency in pricing; in today's marketplace, consumers have ready access to competing quotes from a host of insurers. Insurers seek to increase their market share by refining their risk classification systems, thus achieving a better approximation of the products' prices and enabling cream-skimming underwriting strategies ("cream-skimming" is a phrase used when the insurer underwrites only the best risks). Surveys (e.g., [Earnix \(2013\)](#)) indicate that pricing is the most common use of analytics among insurers.

*Underwriting*, the process of classifying risks into homogeneous categories and assigning policyholders to these categories, lies at the core of ratemaking. Policyholders within a class (category) have similar risk profiles and so are charged the same insurance price. This is the concept of an actuarially fair premium; it is fair to charge different rates to policyholders only if they can be separated by identifiable risk factors. An early article, *Two Studies in Automobile Insurance Ratemaking* ([Bailey and LeRoy, 1960](#)), provided a catalyst to the acceptance of analytic methods in the insurance industry. This paper addresses the problem of classification ratemaking. It describes an example of automobile insurance that has five use classes cross-classified with four merit rating classes. At that time, the contribution to premiums for use and merit rating classes were determined independently of each other. Thinking about the interacting effects of different classification variables is a more difficult problem.

When the risk is initially obtained, the insurer's obligations can be managed by imposing contract parameters that modify contract payouts. Chapter 4 describes common modifications including coinsurance, deductibles and policy upper limits.

### 1.2.2 Renewing Insurance

Insurance is a type of financial service and, like many service contracts, insurance coverage is often agreed upon for a limited time period at which time coverage commitments are complete. Particularly for general insurance, the need for coverage continues and so efforts are made to issue a new contract providing similar coverage when the existing contract comes to the end of its term. This is called *policy renewal*. Renewal issues can also arise in life insur-

ance, e.g., term (temporary) life insurance. At the same time other contracts, such as life annuities, terminate upon the insured's death and so issues of renewability are irrelevant.

In the absence of legal restrictions, at renewal the insurer has the opportunity to:

- accept or decline to underwrite the risk; and
- determine a new premium, possibly in conjunction with a new classification of the risk.

Risk classification and rating at renewal is based on two types of information. First, at the initial stage, the insurer has available many rating variables upon which decisions can be made. Many variables are not likely to change, e.g., sex, whereas others are likely to change, e.g., age, and still others may or may not change, e.g., credit score. Second, unlike the initial stage, at renewal the insurer has available a history of policyholder's loss experience, and this history can provide insights into the policyholder that are not available from rating variables. Modifying premiums with claims history is known as *experience rating*, also sometimes referred to as *merit rating*.

Experience rating methods are either applied retrospectively or prospectively. With retrospective methods, a refund of a portion of the premium is provided to the policyholder in the event of favorable (to the insurer) experience. Retrospective premiums are common in life insurance arrangements (where policyholders earn dividends in the U.S., bonuses in the U.K., and profit sharing in Israeli term life coverage). In general insurance, prospective methods are more common, where favorable insured experience is rewarded through a lower renewal premium.

Claims history can provide information about a policyholder's risk appetite. For example, in personal lines it is common to use a variable to indicate whether or not a claim has occurred in the last three years. As another example, in a commercial line such as worker's compensation, one may look to a policyholder's average claim frequency or severity over the last three years. Claims history can reveal information that is otherwise hidden (to the insurer) about the policyholder.

### 1.2.3 Claims and Product Management

In some types of insurance, the process of paying claims for insured events is relatively straightforward. For example, in life insurance, a simple death certificate is all that is needed to pay the benefit amount as provided in the contract. However, in non-life areas such as property and casualty insurance, the process can be much more complex. Think about a relatively simple insured event

such as an automobile accident. Here, it is often required to determine which party is at fault and then one needs to assess damage to all of the vehicles and people involved in the incident, both insured and non-insured. Further, the expenses incurred in assessing the damages must be assessed, and so forth. The process of determining coverage, legal liability, and settling claims is known as claims adjustment.

Insurance managers sometimes use the phrase claims leakage to mean dollars lost through claims management inefficiencies. There are many ways in which analytics can help manage the claims process, c.f., [Gorman and Swenson \(2013\)](#). Historically, the most important has been fraud detection. The claim adjusting process involves reducing information asymmetry (the claimant knows what happened; the company knows some of what happened). Mitigating fraud is an important part of the claims management process.

Fraud detection is only one aspect of managing claims. More broadly, one can think about claims management as consisting of the following components:

- **Claims triaging.** Just as in the medical world, early identification and appropriate handling of high cost claims (patients, in the medical world), can lead to dramatic savings. For example, in workers compensation, insurers look to achieve early identification of those claims that run the risk of high medical costs and a long payout period. Early intervention into these cases could give insurers more control over the handling of the claim, the medical treatment, and the overall costs with an earlier return-to-work.
- **Claims processing.** The goal is to use analytics to identify routine situations that are anticipated to have small payouts. More complex situations may require more experienced adjusters and legal assistance to appropriately handle claims with high potential payouts.
- **Adjustment decisions.** Once a complex claim has been identified and assigned to an adjuster, analytic driven routines can be established to aid subsequent decision-making processes. Such processes can also be helpful for adjusters in developing case reserves, an estimate of the insurer's future liability. This is an important input to the insurer's loss reserves, described in Section [1.2.4](#).

In addition to the insured's reimbursement for losses, the insurer also needs to be concerned with another source of revenue outflow, expenses. Loss adjustment expenses are part of an insurer's cost of managing claims. Analytics can be used to reduce expenses directly related to claims handling (allocated) as well as general staff time for overseeing the claims processes (unallocated). The insurance industry has high operating costs relative to other portions of the financial services sectors.

In addition to claims payments, there are many other ways in which insurers use data to manage their products. We have already discussed the need for analytics in underwriting, that is, risk classification at the initial acquisition and renewal stages. Insurers are also interested in which policyholders elect to renew their contracts and, as with other products, monitor customer loyalty.

Analytics can also be used to manage the portfolio, or collection, of risks that an insurer has acquired. As described in Chapter ??, after the contract has been agreed upon with an insured, the insurer may still modify its net obligation by entering into a reinsurance agreement. This type of agreement is with a reinsurer, an insurer of an insurer. It is common for insurance companies to purchase insurance on its portfolio of risks to gain protection from unusual events, just as people and other companies do.

#### 1.2.4 Loss Reserving

An important feature that distinguishes insurance from other sectors of the economy is the timing of the exchange of considerations. In manufacturing, payments for goods are typically made at the time of a transaction. In contrast, for insurance, money received from a customer occurs in advance of benefits or services; these are rendered at a later date if the insured event occurs. This leads to the need to hold a reservoir of wealth to meet future obligations in respect to obligations made, and to gain the trust of the insureds that the company will be able to fulfill its commitments. The size of this reservoir of wealth, and the importance of ensuring its adequacy, is a major concern for the insurance industry.

Setting aside money for unpaid claims is known as loss reserving; in some jurisdictions, reserves are also known as *technical provisions*. We saw in Figure 1.1 several times at which a company summarizes its financial position; these times are known as valuation dates. Claims that arise prior to valuation dates have either been paid, are in the process of being paid, or are about to be paid; claims in the future of these valuation dates are unknown. A company must estimate these outstanding liabilities when determining its financial strength. Accurately determining loss reserves is important to insurers for many reasons.

1. Loss reserves represent an anticipated claim that the insurer owes its customers. Under-reserving may result in a failure to meet claim liabilities. Conversely, an insurer with excessive reserves may present a conservative estimate of surplus and thus portray a weaker financial position than it truly has.
2. Reserves provide an estimate for the unpaid cost of insurance that can be used for pricing contracts.

3. Loss reserving is required by laws and regulations. The public has a strong interest in the financial strength and solvency of insurers.
4. In addition to regulators, other stakeholders such as insurance company management, investors, and customers make decisions that depend on company loss reserves. Whereas regulators and customers appreciate conservative estimates of unpaid claims, managers and investors seek more unbiased estimates to represent the true financial health of the company.

Loss reserving is a topic where there are substantive differences between life and general (also known as property and casualty, or non-life) insurance. In life insurance, the severity (amount of loss) is often not a source of uncertainty as payouts are specified in the contract. The frequency, driven by mortality of the insured, is a concern. However, because of the lengthy time for settlement of life insurance contracts, the time value of money uncertainty as measured from issue to date of payment can dominate frequency concerns. For example, for an insured who purchases a life contract at age 20, it would not be unusual for the contract to still be open in 60 years time, when the insured celebrates his or her 80th birthday. See, for example, [Bowers et al. \(1986\)](#) or [Dickson et al. \(2013\)](#) for introductions to reserving for life insurance. In contrast, for most lines of non-life business, severity is a major source of uncertainty and contract durations tend to be shorter.

---

### **1.3 Case Study: Wisconsin Property Fund**

---

In this section, we use the Wisconsin Property Fund as a case study. You learn how to:

- Describe how data generating events can produce data of interest to insurance analysts.
  - Produce relevant summary statistics for each variable.
  - Describe how these summary statistics can be used in each of the major operational areas of an insurance company.
- 

Let us illustrate the kind of data under consideration and the goals that we wish to achieve by examining the Local Government Property Insurance Fund (LGPIF), an insurance pool administered by the Wisconsin Office of the Insurance Commissioner. The LGPIF was established to provide property insur-

ance for local government entities that include counties, cities, towns, villages, school districts, and library boards. The fund insures local government property such as government buildings, schools, libraries, and motor vehicles. It covers all property losses except those resulting from flood, earthquake, wear and tear, extremes in temperature, mold, war, nuclear reactions, and embezzlement or theft by an employee.

The fund covers over a thousand local government entities who pay approximately 25 million dollars in premiums each year and receive insurance coverage of about 75 billion. State government buildings are not covered; the LGPIF is for local government entities that have separate budgetary responsibilities and who need insurance to moderate the budget effects of uncertain insurable events. Coverage for local government property has been made available by the State of Wisconsin since 1911, thus providing a wealth of historical data.

In this illustration, we restrict consideration to claims from coverage of building and contents; we do not consider claims from motor vehicles and specialized equipment owned by local entities (such as snow plowing machines). We also consider only claims that are closed, with obligations fully met.

### 1.3.1 Fund Claims Variables: Frequency and Severity

At a fundamental level, insurance companies accept premiums in exchange for promises to compensate a policyholder upon the occurrence of an insured event. Indemnification is the compensation provided by the insurer for incurred hurt, loss, or damage that is covered by the policy. This compensation is also known as a *claim*. The extent of the payout, known as the *severity*, is a key financial expenditure for an insurer.

In terms of money outgo to customers, an insurer is indifferent to having ten claims of 100 when compared to one claim of 1,000. Nonetheless, it is common for insurers to study how often claims arise, known as the *frequency* of claims. The frequency is important for expenses, but it also influences contractual parameters (such as deductibles and policy limits that are described later) that are written to limit amounts paid for each occurrence of an insured event. Frequency is routinely monitored by insurance regulators and can be a key driver in the overall indemnification obligation of the insurer. We shall consider the frequency and severity as the two main claim variables that we wish to understand, model, and manage.

To illustrate, in 2010 there were 1,110 policyholders in the property fund who experienced a total of 1,377 claims. Table 1.1 shows the distribution. Almost two-thirds (0.637) of the policyholders did not have any claims and an additional 18.8% had only one claim. The remaining 17.5% ( $=1 - 0.637 - 0.188 = 0.175$ ) had two or more claims.

TABLE 1.1: 2010 Claims Frequency Distribution

Number	0	1	2	3	4	5	6	7	8	9 or more	Sum
Policies	707	209	86	40	18	12	9	4	6	19	1110
Claims	0	209	172	120	72	60	54	28	48	614	1377
Proportion	0.637	0.188	0.077	0.036	0.016	0.011	0.008	0.004	0.005	0.017	1

TABLE 1.2: 2010 Average Severity Distribution

Minimum	First Quartile	Median	Mean	Third Quartile	Maximum
167	2,226	4,951	56,332	11,900	12,922,218

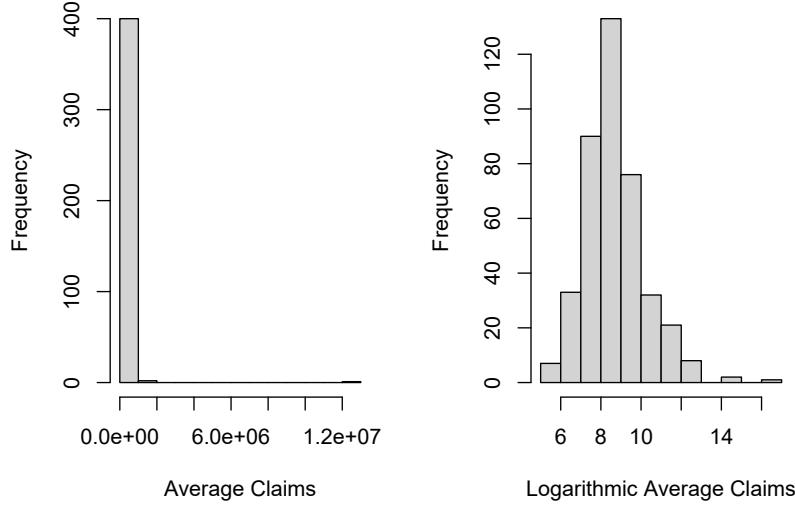
0.188) had more than one claim; the policyholder with the highest number recorded 239 claims. The average number of claims for this sample was 1.24 ( $=1377/1110$ ).

For the severity distribution, a common approach is to examine the distribution of the sample of 1,377 claims. However, another common approach is to examine the distribution of the average claims of those policyholders with claims. In our 2010 sample, there were 403 ( $=1110-707$ ) such policyholders. For 209 of these policyholders with one claim, the average claim equals the only claim they experienced. For the policyholder with highest frequency, the average claim is an average over 239 separately reported claim events.

Table 1.2 summarizes the sample distribution of average severities from the 403 policyholders who made a claim; it shows that the average claim amount was 56,330 (all amounts are in U.S. Dollars). However, the average gives only a limited look at the distribution. More information can be gleaned from the summary statistics which show a very large claim in the amount of 12,920,000. Figure 1.2 provides further information about the distribution of sample claims, showing a distribution that is dominated by this single large claim so that the histogram is not very helpful. Even when removing the large claim, you will find a distribution that is skewed to the right. A generally accepted technique is to work with claims in logarithmic units especially for graphical purposes; the corresponding figure in the right-hand panel is much easier to interpret.

### 1.3.2 Fund Rating Variables

Developing models to represent and manage the two outcome variables, frequency and severity, is the focus of the early chapters of this text. However, when actuaries and other financial analysts use those models, they do so in the context of external variables. In general statistical terminology, one might call these explanatory or predictor variables; there are many other names in



**FIGURE 1.2: Distribution of Positive Average Severities**

statistics, economics, psychology, and other disciplines. Because of our insurance focus, we call them rating variables as they are useful in setting insurance rates and premiums.

We earlier considered observations from a sample of 1,110 policyholders which may seem like a lot. However, as we will see in our forthcoming applications, because of the preponderance of zeros and the skewed nature of claims, actuaries typically yearn for more data. One common approach that we adopt here is to examine outcomes from multiple years, thus increasing the sample size. We will discuss the strengths and limitations of this strategy later but, at this juncture, we just wish to show the reader how it works.

Specifically, Table 1.3 shows that we now consider policies over five years of data, 2006, ..., 2010, inclusive. The data begins in 2006 because there was a shift in claim coding in 2005 so that comparisons with earlier years are not helpful. To mitigate the effect of open claims, we consider policy years prior to 2011. An open claim means that not all of the obligations for the claim are known at the time of the analysis; for some claims, such an injury to a person in an auto accident or in the workplace, it can take years before costs are fully known.

Table 1.3 shows that the average claim varies over time, especially with the

TABLE 1.3: Claims Summary by Policyholder

Year	Average Frequency	Average Severity	Average	Number of Policyholders
2006	0.951	9,695	32,498,186	1,154
2007	1.167	6,544	35,275,949	1,138
2008	0.974	5,311	37,267,485	1,125
2009	1.219	4,572	40,355,382	1,112
2010	1.241	20,452	41,242,070	1,110

TABLE 1.4: Summary of Claim Frequency and Severity, Deductibles, and Coverages

	Minimum	Median	Average	Maximum
Claim Frequency	0	0	1.109	263
Claim Severity	0	0	9,292	12,922,218
Deductible	500	1,000	3,365	100,000
Coverage (000's)	8.937	11,354	37,281	2,444,797

high 2010 value (that we saw was due to a single large claim)<sup>1</sup>. The total number of policyholders is steadily declining and, conversely, the coverage is steadily increasing. The coverage variable is the amount of coverage of the property and contents. Roughly, you can think of it as the maximum possible payout of the insurer. For our immediate purposes, the coverage is our first rating variable. Other things being equal, we would expect that policyholders with larger coverage have larger claims. We will make this vague idea much more precise as we proceed, and also justify this expectation with data.

For a different look at the 2006-2010 data, Table 1.4 summarizes the distribution of our two outcomes, frequency and claims amount. In each case, the average exceeds the median, suggesting that the two distributions are right-skewed. In addition, the table summarizes our continuous rating variables, coverage and deductible amount. The table also suggests that these variables also have right-skewed distributions.

Table 1.5 describes the rating variables considered in this chapter. Hopefully, these are variables that you think might naturally be related to claims outcomes. You can learn more about them in [Frees et al. \(2016\)](#). To handle the

---

<sup>1</sup>Note that the average severity in Table 1.3 differs from that reported in Table 1.2. This is because the former includes policyholders with zero claims where as the latter does not. This is an important distinction that we will address in later portions of the text.

skewness, we henceforth focus on logarithmic transformations of coverage and deductibles.

Table 1.5. Description of Rating Variables

<i>Variable</i>	<i>Description</i>
EntityType	Categorical variable that is one of six types: (Village, City, County, Misc, School, or Town)
LnCoverage	Total building and content coverage, in logarithmic millions of dollars
LnDeduct	Deductible, in logarithmic dollars
AlarmCredit	Categorical variable that is one of four types: (0, 5, 10, or 15) for automatic smoke alarms in main rooms
NoClaimCredit	Binary variable to indicate no claims in the past two years
Fire5	Binary variable to indicate the fire class is below 5 (The range of fire class is 1 to 10)

For the *alarm credit* variable, a zero means that no automatic smoke alarms exist in any of the main rooms. In the same way, a 5 means they exist in some of the main rooms and a 10 means they exist in all of the main rooms. At the 15 level, facilities are monitored on a 24 hours per day, 7 days per week basis by a police, fire, or security company. A *fire rating* is a similar type of score. It reflects how prepared a community and area is for fires. While it mainly focuses on the local fire departments and water supply, there are other factors that contribute to an area's score. This rating is used to determine how likely it is for a fire to do severe damage before help arrives with 1 being the best and 10 the worst.

To get a sense of the relationship between the non-continuous rating variables and claims, Table 1.6 relates the claims outcomes to these categorical variables. Table 1.6 suggests substantial variation in the claim frequency and average severity of the claims by entity type. It also demonstrates higher frequency and severity for the *Fire5* variable and the reverse for the *NoClaimCredit* variable. The relationship for the *Fire5* variable is counter-intuitive in that one would expect lower claim amounts for those policyholders in areas with better public protection (when the protection code is five or less). Naturally, there are other variables that influence this relationship. We will see that these background variables are accounted for in the subsequent multivariate regression analysis, which yields an intuitive, appealing (negative) sign for the *Fire5* variable.

Tables 1.7 and 1.8 show the claims experience by alarm credit. It underscores the difficulty of examining variables individually. For example, when looking at the experience for all entities, we see that policyholders with no alarm credit have on average lower frequency and severity than policyholders with the highest (15%, with 24/7 monitoring by a fire station or security company) alarm credit. In particular, when we look at the entity type School, the frequency

TABLE 1.6: Claims Summary by Entity Type, Fire Class, and No Claim Credit

	Number of Policies	Claim Frequency	Average Severity
Village	1,341	0.452	10,645
City	793	1.941	16,924
County	328	4.899	15,453
Misc	609	0.186	43,036
School	1,597	1.434	64,346
Town	971	0.103	19,831
Fire5-No	2,508	0.502	13,935
Fire5-Yes	3,131	1.596	41,421
NoClaimCredit-No	3,786	1.501	31,365
NoClaimCredit-Yes	1,853	0.31	30,499
Total	5,639	1.109	31,206

is 0.422 and the severity 25,523 for no alarm credit, whereas for the highest alarm level it is 2.008 and 85,140, respectively. This may simply imply that entities with more claims are the ones that are likely to have an alarm system. Summary tables do not examine multivariate effects; for example, Table 1.6 ignores the effect of size (as we measure through coverage amounts) that affect claims.

We will learn more about modeling count data in the Chapter 3 and about severity data in Chapters 4 and 7.

### 1.3.3 Fund Operations

We have now seen distributions of the Fund's two outcome variables: a count variable for the number of claims, and a continuous variable for the claims amount. We have also introduced a continuous rating variable (logarithmic coverage); a discrete quantitative variable (logarithmic deductibles); two binary rating variables (no claims credit and fire class); and two categorical rating variables (entity type and alarm credit). Subsequent chapters will explain how to analyze and model the distribution of these variables and their relationships. Before getting into these technical details, let us first think about where we want to go. General insurance company functional areas are described in Section 1.2; we now consider how these areas might apply in the context of the property fund.

**TABLE 1.7: Claims Summary by Entity Type and Alarm Credit (AC) Categories 0 and 5**

	AC0 Claim Fre- quency	AC0 Avg. Severity	AC0 Num. Policies	AC5 Claim Fre- quency	AC5 Avg. Severity	AC5 Num. Policies
Village	0.326	11,078	829	0.278	8,086	54
City	0.893	7,576	244	2.077	4,150	13
County	2.14	16,013	50	0	0	1
Misc	0.117	15,122	386	0.278	13,064	18
School	0.422	25,523	294	0.41	14,575	122
Town	0.083	25,257	808	0.194	3,937	31
Total	0.318	15,118	2611	0.431	10,762	239

**TABLE 1.8: Claims Summary by Entity Type and Alarm Credit (AC) Categories 10 and 15**

	AC10 Claim Fre- quency	AC10 Avg. Severity	AC10 Num. Policies	AC15 Claim Fre- quency	AC15 Avg. Severity	AC15 Num. Policies
Village	0.5	8,792	50	0.725	10,544	408
City	1.258	8,625	31	2.485	20,470	505
County	2.125	11,688	8	5.513	15,476	269
Misc	0.077	3,923	26	0.341	87,021	179
School	0.488	11,597	168	2.008	85,140	1013
Town	0.091	2,338	44	0.261	9,490	88
Total	0.517	10,194	327	2.093	41,458	2462

### Initiating Insurance

Because this is a government sponsored fund, we do not have to worry about selecting good or avoiding poor risks; the fund is not allowed to deny a coverage application from a qualified local government entity. If we do not have to underwrite, what about how much to charge?

We might look at the most recent experience in 2010, where the total fund claims were approximately 28.16 million USD ( $= 1377 \text{ claims} \times 20452 \text{ average severity}$ ). Dividing that among 1,110 policyholders, that suggests a rate of 24,370 ( $\approx 28,160,000/1110$ ). However, 2010 was a bad year; using the same method, our premium would be much lower based on 2009 data. This swing in premiums would defeat the primary purpose of the fund, to allow for a steady charge that local property managers could utilize in their budgets.

Having a single price for all policyholders is nice but hardly seems fair. For example, Table 1.6 suggests that schools have higher aggregate claims than other entities and so should pay more. However, simply doing the calculation on an entity by entity basis is not right either. For example, we saw in Tables 1.7 and 1.8 that had we used this strategy, entities with a 15% alarm credit (for good behavior, having top alarm systems) would actually wind up paying more.

So, we have the data for thinking about the appropriate rates to charge but need to dig deeper into the analysis. We will explore this topic further in Chapter 10 on *premium calculation fundamentals*. Selecting appropriate risks is introduced in Chapter ?? on *risk classification*.

### Renewing Insurance

Although property insurance is typically a one-year contract, Table 1.3 suggests that policyholders tend to renew; this is typical of general insurance. For renewing policyholders, in addition to their rating variables we have their claims history and this claims history can be a good predictor of future claims. For example, Table 1.6 shows that policyholders without a claim in the last two years had much lower claim frequencies than those with at least one accident (0.310 compared to 1.501); a lower predicted frequency typically results in a lower premium. This is why it is common for insurers to use variables such as `NoClaimCredit` in their rating. We will explore this topic further in Chapters ?? and ?? on *experience rating*.

### Claims Management

Of course, the main story line of the 2010 experience was the large claim of over 12 million USD, nearly half the amount of claims for that year. Are there ways

that this could have been prevented or mitigated? Are there ways for the fund to purchase protection against such large unusual events? Another unusual feature of the 2010 experience noted earlier was the very large frequency of claims (239) for one policyholder. Given that there were only 1,377 claims that year, this means that a single policyholder had 17.4 % of the claims. These extreme features of the data suggests opportunities for managing claims, the subject of Chapter ??.

### Loss Reserving

In our case study, we look only at the one year outcomes of closed claims (the opposite of open). However, like many lines of insurance, obligations from insured events to buildings such as fire, hail, and the like, are not known immediately and may develop over time. Other lines of business, including those where there are injuries to people, take much longer to develop. Chapter ?? introduces this concern and *loss reserving*, the discipline of determining how much the insurance company should retain to meet its obligations.

---

## 1.4 Exercises

These exercises ask you to work with data using statistical software, such as R code. If you would like some practice with R code, please visit the [first chapter of a Short Course on Loss Data Analytics](#). As another method of learning, you can also get practice executing 'R' code at our [Online Version R Code Site](#).

**Exercise 1.1. Corporate Travel.** Universities purchase corporate travel policies to cover employees and students traveling on official university business for a wide variety of accidents and incidents while away from the campus or primary workplace. This broad coverage includes medical care and evacuation, loss of personal property, extraction for political and weather related reasons, and more. These data represent experience from the Australian National University (ANU) and additional details can be found in [ANU's corporate travel policy](#). You can also learn more about this line of business from ANU's insurer, [Chubb Travel](#). The data provided are maintained by the insurer, Chubb, and were accessed on 29 July 2022. You can retrieve the data by going to Appendix Section ??.

a. *Claim Frequency.* The travel data history is long and stable. This coverage began on 1 November 2006. Table 1.9 shows the count of claims for years 2015-2019, inclusive. Produce a comparable table of claims frequency for the entire period. Comment on the unusual frequency surrounding the COVID pandemic.

TABLE 1.9: 2015-2019 Travel Claims Frequency

2015	2016	2017	2018	2019
158	154	139	205	274

b. *Adjust for Zero Claims.* From this data set, there are 2107 incurred claims. Of these claims, there are 269 zeros and an additional 3 claims where the incurred claim is less than 10. We omit these claims in our analysis. Reproduce your part (a) analysis by omitting incurred claims less than 10.

c. *Loss Distributions over Time.* There are 1835 incurred losses in the dataset with all available years (yet omitting claims less than 10). Figure 1.3 shows that the distribution of incurred losses is stable over the period 2015-2019, inclusive. Produce a comparable figure for the entire period.

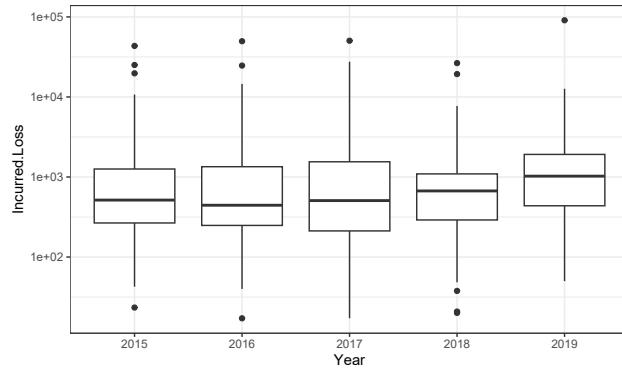


FIGURE 1.3: Distribution of Travel Losses by Year

d. *Summary Statistics.* In addition to graphs, it can be helpful to display several summary statistics. For the five year period 2015-2019, produce a set of summary statistics.

**Exercise 1.2. Group Personal Accident.** Group personal accident insurance offers financial protection in case of injury or death resulting from an incident that occurs on the job. Group personal accident offers insurance coverage and liability insurance protection against accidental death or injury. The insurance covers students and ANU's voluntary workers; ANU workers are covered through another system known as "workers' compensation."

Several limits apply including 1,000,000 for the period of insurance, 600,000 for non-scheduled flights, and others. These limits were not reached in the data we consider. For this coverage, there is a "7 day excess" for weekly benefits but

TABLE 1.10: 2015-2019 Group Personal Accident Claims Frequency

2015	2016	2017	2018	2019
4	7	16	11	9

none for general benefits. The database documentation provided to us, and the data we provide, do not indicate whether the excess has been triggered; we have only paid claims. Because of the relatively small size of this class of insurance, we ignore the effects of deductibles for this line.

The data provided to us are maintained by the insurer, Chubb. These data began in underwriting year 2007 and were accessed on 29 July 2022. You can retrieve the data by going to Appendix Section ??.

a. *Claim Frequency.* From this data set, there are 148 incurred claims. Of these claims, there are 35 zeros and an additional 0 claims where the incurred claim is less than 10. We omit these claims in our analysis. Table 1.10 shows the count of claims for years 2015-2019, inclusive. Produce a comparable table of claims frequency for the entire period, omitting claims that are less than 10.

b. *Skewness of Claims Severity Distribution.* The left-hand panel of Figure 1.4 shows a histogram of incurred claims that reveals the right-skewed nature of this distribution. The right-hand panel shows the same claims but on the log (base 10) scale; this plot demonstrates that the log transform can symmetrize a distribution. These plots are for the 2015-2019 data. Reproduce this work, using incurred claims for all available years (still omitting those less than 10).

c. *Summary Statistics.* Produce summary statistics for both claims and log claims using all available years (still omitting those less than 10). Comment on the relationship between the mean and the median for both claims and log claims, relating this to the symmetry of the distributions observed in part (b).

d. *Loss Distributions over Time.* There are 112 incurred losses. Figure 1.5 indicates that the incurred losses are stable over the period 2015-2019, inclusive. Produce a comparable figure for the entire period and comment on the stability of the distribution.

---

**Exercise 1.3. Motor Vehicle.** This policy covers ANU's vehicles including cars, vans, utilities, and motorcycles. There are two parts to this coverage, one for comprehensive damage to the insured vehicles and a second for legal liability. The comprehensive coverage for loss or damage is essentially limited by the market value of the insured vehicle. For legal liability, there is a \$50 Million upper limit for all claims arising from the one accident or series of

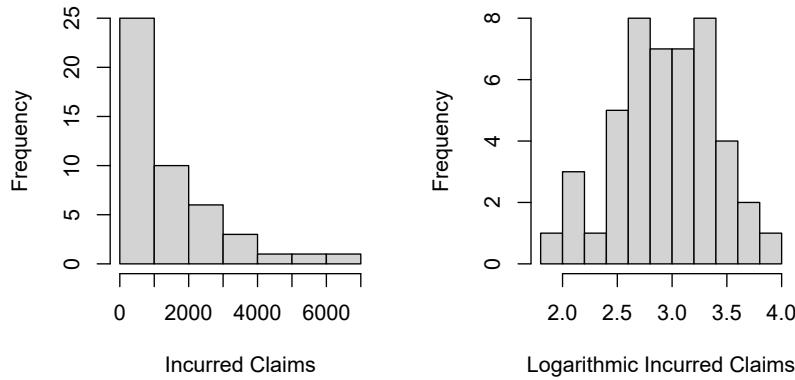


FIGURE 1.4: Distribution of Incurred Claims 2015-2019

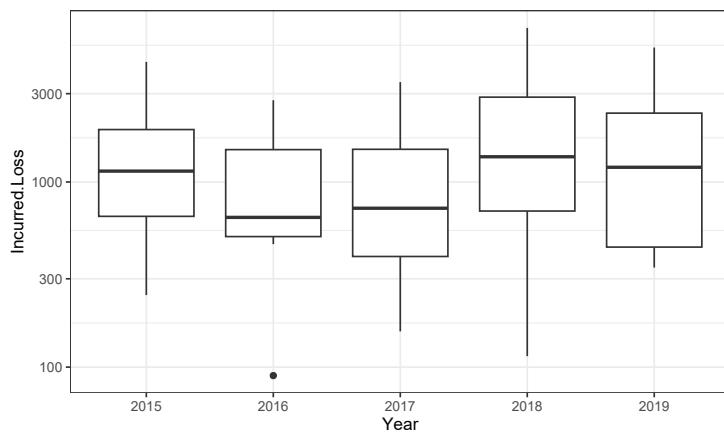


FIGURE 1.5: Distribution of Group Personal Accident Losses by Year

accidents resulting from the one original cause. There is also another upper limit (that is lower than 50 million) when the vehicle is used for transportation of dangerous goods.

The data available contain the amount paid by the insurer (Vero Insurance Limited) which is the focus of our initial analysis. In addition, the data also contains a deductible (called an “excess” in the data file) that we explore in later parts.

The data provided to us are maintained by the insurer, Vero Insurance Limited. These data began in underwriting year 2012 and were accessed on 8 August 2022. You can retrieve the data by going to Appendix Section ??.

*a. Adjust for Zeros.* From this data set, check that:

- there are 318 incurred claims.
- Of these claims, there are 50 zeros and
- an additional 0 claims where the incurred claim is less than 10.

Remove these claims in your analysis, so that there are 268 incurred claims.

*b. Claim Frequency.* Produce a table that shows the count of claims for the entire period.

*c. Loss Distributions over Time.* Produce a figure that shows the distribution of motor vehicle paid amounts over time and comment on the stability of the distribution.

*d. Year 2019.* In your analysis from the prior steps, you may have noticed the unusual aspects of year 2019. In that year, ANU suffered extensive damage from a hailstorm that increased the frequency of claims as well as the severity. Produce a histogram of paid claims for that year.

*e. Deductibles.* For each event, or series of events arising from the one originating cause, ANU bears the amount of the excess in respect of each and every insured vehicle, unless stated otherwise. The standard deductible (or excess) in the dataset is 1000. However, a cursory examination of the dataset shows tremendous variation by vehicle and over time. Replicate Table 1.11 that shows, for each year, the number of claims with zero excess, positive excess less than 1000, an excess equal to 1000, and an excess greater than 1000.

(**Deductibles.** We recommend that motivated readers extend our analysis to account for this deductible in both the severity and frequency.)

---

TABLE 1.11: Motor Vehicle Excess by Year

UW.Year	Num 0	Num 0-1000	Num = 1000	Num >1000	Total
2011	1	1	7	0	9
2012	1	2	13	0	16
2013	4	1	22	0	27
2014	0	0	11	0	11
2015	1	1	14	0	16
2016	6	1	19	0	26
2017	16	0	4	1	21
2018	19	0	1	0	20
2019	99	0	6	0	105
2020	5	0	0	0	5
2021	10	0	0	0	10

## 1.5 Further Resources and Contributors

If you would like additional practice with R coding, please visit our companion [LDA Short Course](#). In particular, see the [Introduction to Loss Data Analytics Chapter](#).

### Contributor

- **Edward (Jed) Frees**, University of Wisconsin-Madison and Australian National University, is the principal author of the initial version and second edition of this chapter. Email: [jfrees@bus.wisc.edu](mailto:jfrees@bus.wisc.edu) for chapter comments and suggested improvements.
- Chapter reviewers include: Yair Babad, Chunsheng Ban, Aaron Bruhn, Gordon Enderle, Hirokazu (Iwahiro) Iwasawa, Dalia Khalil, Bell Ouelega, Michelle Xia.

This book introduces loss data analytic tools that are most relevant to actuaries and other financial risk analysts. We have also introduced you to many new insurance terms; more terms can be found at the [NAIC Glossary \(2018\)](#).

This work is licensed under a Creative Commons Attribution 4.0 International License.



# 2

---

## *Introduction to Data Analytics*

---

*Chapter Preview.* This introduction focuses on data analytics concepts relevant to insurance activities. As data analytics is used across various fields with different terminologies, we start in Section 2.1 by describing the basic ingredients or elements of data analytics. Then, Section 2.2 outlines a process an analyst can use to analyze insurance data. Many fields emphasize the development of data analytics with a focus on multiple variables, or “big” data. However, this often comes at the cost of excluding consideration of a single variable. So, Section 2.3 introduces an approach we call “single variable analytics,” which includes a description of variable types, exploratory versus confirmatory analysis, and elements of model construction and selection, all of which can be done in the context of a single variable. Building on this, Section 2.4 explores the roles of supervised and unsupervised learning, which require the presence of many variables.

The final section of this chapter, Section 2.5, offers a broader introduction to data considerations beyond the scope of this book, intended for budding analysts who want to use this chapter to build a foundation for further studies in data analytics. Additionally, the technical supplements introduce other standard ingredients of data analytics, such as principal components, cluster analysis, and tree-based regression models. While these topics are not necessary for this book, they are important in a broader analytics context.

---

### **2.1 Elements of Data Analytics**

---

In this section, you learn how to describe the essential ingredients of data analytics

- consisting of several key concepts, and
  - two fundamental approaches, data and algorithmic modeling.
-

**Data analysis** involves inspecting, cleansing, transforming, and modeling data to discover useful information to suggest conclusions and make decisions. Data analysis has a long history. In 1962, statistician John Tukey defined data analysis as:

procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

— (Tukey, 1962)

### 2.1.1 Key Data Analytic Concepts

Underpinning the elements of data analytics are the following key concepts:

- **Data Driven.** As described in Section 1.1.2, the conclusions and decisions made through a data analytic process depend heavily on data inputs. In comparison, econometricians have long recognized the difference between a data-driven model and a structural model, the latter being one that represents an explicit interplay between economic theory and stochastic models, [Goldberger \(1972\)](#).
- **EDA** - exploratory data analysis - and **CDA** - confirmatory data analysis. Although some techniques overlap, e.g., taking the average of a dataset, these two approaches to analyzing data have different purposes. The purpose of EDA is to reveal aspects or patterns in the data without reference to any particular model. In contrast, CDA techniques use data to substantiate, or confirm, aspects or patterns in a model. See Section 2.3.2 for further discussions.
- **Estimation and Prediction.** Recall the traditional triad of statistical inference: hypothesis testing, parameter estimation, and prediction. Medical statisticians test the efficacy of a new drug and econometricians estimate parameters of an economic relationship. In insurance, one also uses hypothesis testing and parameter estimation. Moreover, predictions of yet to be realized random outcomes are critical for financial risk management (e.g., pricing) of existing risks in future periods, as well as not yet observed risks in a current period, cf. [Frees \(2015\)](#).
- **Model Complexity, Parsimony, and Interpretability.** A model is a mathematical representation of reality that, in statistics, is calibrated using a data set. One concern is the *complexity* of the model where the complexity may involve the number of parameters used to define the model, the number of variables upon which it relies, and the intricacies of relationships among the parameters and variables. As a rule of thumb, we will see that the more

complex is the model, the better it fares in fitting a set of data (and hence at estimation) but the worse it fares in predicting new outcomes. Other things being equal, a model with fewer parameters is said to be *parsimonious* and hence less complex. Moreover, a parsimonious model is typically easier to interpret than a comparable model that is more complex. Complexity hinders our ability to understand the inner workings of a model, its interpretability, and will be a key ingredient in our comparisons of data versus algorithmic models in Section 2.1.2.

- **Parametric and Nonparametric** models. Many models, including stochastic distributions, are known with the exception of a limited number of quantities known as parameters. For example, the mean and variance are parameters that determine a normal distribution. In contrast, other models may not rely on parameters; these are simply known as *nonparametric* models. Naturally, there is also a host of models that rely on parameters for some parts of the distribution and are distribution-free for other portions; these are referred to as *semi-parametric* models. Parametric and nonparametric approaches have different strengths and limitations; neither is strictly better than the other. We start the discussion in Section 2.3.3 to explain under what circumstances you might prefer one approach to another.
- **Robustness** means that a model, test, or procedure is resistant to unanticipated deviations in model assumptions or the data used to calibrate the model. When interpreting findings, it is natural to ask questions about how the results react to changes in assumptions or data, that is, the robustness of the results.
- **Computational Statistics.** Historically, statistical modeling relied extensively on summary statistics that were not only easy to interpret but also easy to compute. With modern-day computing power, definitions of “easy to compute” have altered drastically paving the way for measures that were once deemed far too computationally intensive to be of practical use. Moreover, ideas of subsampling and resampling data (e.g., through cross-validation and bootstrapping) have introduced new methods for understanding statistical sampling errors and a model’s predictive capabilities.
- **Big Data.** This is about the process of using special methods and tools that can extract information rapidly from massive data. Examples of big data include text documents, videos, and audio files that are also known as *unstructured* data. Table 2.1 summarizes new types of data sources that lead to new data. As part of the analytics trends, different types of algorithms lead to new software for handling new types of data. See Section 2.5.4 for further discussions.

Table 2.1. Analytic Trends (from [Frees and Gao \(2019\)](#))

Data Sources	Algorithms
Mobile devices	Statistical learning
Auto telematics	Artificial intelligence
Home sensors (Internet of Things)	Structural models
Drones, micro satellites	
Data	Software
Big data (text, speech, image, video)	Text analysis, semantics
Behavioral data (including social media)	Voice recognition
Credit, trading, financial data	Image recognition Video recognition

*Source :* Stephen Mildenhall, Personal Communication

### 2.1.2 Data versus Algorithmic Modeling

There are two cultures for the use of statistical modeling to reach conclusions from data: the data modeling culture and the algorithmic modeling culture. In the data modeling culture, data are assumed to be generated by a given stochastic model. In the algorithmic modeling culture, the data mechanism is treated as unknown and algorithmic models are used.

Data modeling allows statisticians to analyze data and acquire information about the data mechanisms. However, [Breiman \(2001\)](#) argued that the focus on data modeling in the statistical community has led to some side effects such as:

- It produced irrelevant theory and questionable scientific conclusions.
- It kept statisticians from using algorithmic models that might be more suitable.
- It restricted the ability of statisticians to deal with a wide range of problems.

Algorithmic modeling was used by industrial statisticians long time ago. Sadly, the development of algorithmic methods was taken up by communities outside statistics. The goal of algorithmic modeling is predictive accuracy. For some complex prediction problems, data models are not suitable. These prediction problems include voice recognition, image recognition, handwriting recognition, nonlinear time series prediction, and financial market prediction. The theory in algorithmic modeling focuses on the properties of algorithms, such as convergence and predictive accuracy.

---

## 2.2 Data Analysis Process

---

In this section, you learn how to describe the data analysis process as five steps:

- scoping phase,
  - data splitting,
  - model development,
  - validation, and
  - determining implications.
- 

**Table 2.2** outlines common steps used when analyzing data associated with insurance activities.

Table 2.2 Data Analysis Process for Insurance Activities

I. Scoping Phase	II. Data Splitting	III. Model Development	IV. Validation	V. Determine Implications
Use background knowledge and theory to define goals  Prepare, collect, and revise data  EDA Explore the data	Split the data into training and testing portions  Select variables to be used with the candidate model  Evaluate model fit using training data  Use deviations from model fit to improve suggested models	Select a candidate model  Assess each model using the testing portion of the data to determine its predictive capabilities	Repeat Phase III to determine several candidate models	Use knowledge gained from exploring the data, fitting and predicting the models to make data-informed statements about the project goals

### I. Scoping Phase

Scoping, or problem formulation, can be divided into three components:

- **Use background knowledge and theory to define goals.** Insurance activity projects are commonly motivated by business pursuits that have been formulated to be consistent with background knowledge such as market conditions and theory such as a person's attitude towards risk-taking.

- **Prepare, collect, and revise data.** Getting the right data that gives insights into questions at hand is typically the most time-consuming aspect of most projects. Section 2.5 delves more into the devilish details of data structures, quality, cleaning, and so forth.
- **EDA** - Exploring the data, without reference to any particular model, can reveal unsuspected aspects or patterns in the data.

These three components can be performed *iteratively*. For example, a question may suggest collecting certain data types. Then, a preliminary analysis of the data raises additional questions of interest that can lead to seeking more data - this cycle can be repeated many times. Note that this iterative approach differs from the traditional “scientific method” whereby the analyst develops a hypothesis, collects data, and then employs the data to test the hypothesis.

## II. Data Splitting

Although optional, splitting the data into training and testing portions has some important advantages. If the available dataset is sufficiently large, one can split the data into a portion used to calibrate one or more candidate models, the training portion, and another portion that can be used for testing, that is, evaluating the predictive capabilities of the model. The data splitting procedure guards against overfitting a model and emphasizes predictive aspects of a model. For many applications, the splitting is done randomly to mitigate unanticipated sources of bias. For some applications such as insurance, it is common to use data from an earlier time period to predict, or *forecast*, future behavior. For example, with the Section 1.3 Wisconsin Property Fund data, one might use 2006-2010 data for training and 2011 data for assessing predictions.

For large datasets, some analysts prefer to split the data into three portions, one for training (model estimation), one for validation (estimate prediction error for model selection), and one for testing (assessment of the generalization error of the final chosen model), c.f. [Hastie et al. \(2009\)](#) (Chapter 7). In contrast, for moderate and smaller datasets, it is common to use cross-validation techniques where one repeatedly splits the dataset into training and testing portions and then averages results over many applications. These techniques are described further in Chapter 8.

## III. Model Development

The objective of the model development phase is to consider different types of model and provide the best fit for each “candidate” model. As with the scoping phase, developing a model is an iterative procedure.

- **Select a candidate model.** One starts with a model that, from the an-

alyst's perspective, is a likely "candidate" to be the recommended model. Although analysts will focus on familiar models, such as through their past applications of a model or its acceptance in industry, in principle one remains open to all types of models.

- **Select variables to be used with the candidate model.** For simpler situations, only a single outcome, or variable, is of interest. However, many (if not most) situations deal with multivariate outcomes and, as will be seen in Section 2.4, analysts give a great deal of thought as to which variables are considered inputs to a system and which variables can be treated as outcomes.
- **Evaluate model fit on training data.** Given a candidate model based on one or more selected variables, the next step is to calibrate the model based on the training data and evaluate the model fit. Many measures of model fit are available - analysts should focus on those likely to be consistent with the project goals and intended audience of the data analysis process.
- **Use deviations from the model fit to suggest improvements to the candidate model.** When comparing the training data to model fits, it may be that certain patterns are revealed that suggest model improvements. In regression analysis, this tactic is known as *diagnostic checking*.

#### IV. Validation

- **Repeat Phase III to determine several candidate models.** There is a wealth of potential models from which an analyst can choose. Some are parametric, others non-parametric, and some a mixture between the two. Some focus on simplicity such as through linear relationships whereas others are much more complex. And so on. Through repeated applications of the Phase III process, it is customary to narrow the field of candidates down to a handful based on their fit to the training data.
- **Assess each model using the testing portion of the data to determine its predictive capabilities.** With the handful of models that perform the best in the model development phase, one assesses the predictive capabilities of each model. Specifically, each fitted model is used to make predictions with the predicted outcomes compared to the held-out test data. This comparison may also be done using cross-validation. Models are then compared based on their predictive capabilities.

#### V. Determine Implications

The scoping, model development, and validation phases all contribute to making data-informed statements about the project goals. Although most projects result in a single recommended model, each phase has the potential to lend powerful insights.

For data analytic projects associated with insurance activities, it is common to select the model with best predictive capabilities. However, analysts are also mindful of the intended audiences of their analyses, and it is also common to favor models that are simpler and easier to interpret. The relative importance of interpretability very much depends on the project goals. For example, a model devoted to enticing potential customers to view a webpage can be judged more on its predictive capabilities. In contrast, a model that provides the foundations for insurance prices typically undergoes scrutiny by regulators and consumer advocacy groups; here, interpretation plays an important role.

---

## 2.3 Single Variable Analytics

---

In this section, you learn how to describe analytics based on a single variable in terms of

- the type of variable,
  - exploratory versus confirmatory analyses,
  - model construction and
  - model selection.
- 

Rather than starting with multiple variables consisting of inputs and outputs as is common in analytics, in this section we restrict considerations to a single variable. Single variable analytics is motivated by statistical data modeling. Moreover, as will be seen in Chapters 3-8, single variable analytics plays a prominent role in fundamental insurance and risk management applications.

### 2.3.1 Variable Types

This section describes basic variable types traditionally encountered in statistical data analysis. Section 2.5 will provide a framework for more extensive types that include big data.

#### Qualitative Variables

A qualitative, or categorical variable is one for which the measurement denotes membership in a set of groups, or categories. For example, if you were coding in which area of the country an insured resides, you might use 1 for the northern part, 2 for southern, and 3 for everything else. Any analysis of categorical variables should not depend on the labeling of the categories. For example,

instead of using a 1,2,3 for north, south, other, one should arrive at the same set of summary statistics if I used a 2,1,3 coding instead, interchanging north and south.

In contrast, an ordinal variable is a variation of categorical variable for which an ordering exists. For example, with a survey to see how satisfied customers are with our claims servicing department, we might use a five point scale that ranges from 1 meaning dissatisfied to a 5 meaning satisfied. Ordinal variables provide a clear ordering of levels of a variable although the amount of separation between levels is unknown.

A binary variable is a special type of categorical variable where there are only two categories commonly taken to be 0 and 1.

Earlier, in the Section 1.3 case study, we saw in [Table 1.5](#) several examples of qualitative variables. These included the categorical `EntityType` and binary variables `NoClaimCredit` and `Fire5`. We also treated `AlarmCredit` as a categorical variable although some analysts may wish to explore its use as an ordinal variable.

### Quantitative Variables

Unlike a qualitative variable, a quantitative variable is one in which each numerical level is a realization from some scale so that the distance between any two levels of the scale takes on meaning. A continuous variable is one that can take on any value within a finite interval. For example, one could represent a policyholder's age, weight, or income, as continuous variables. In contrast, a discrete variable is one that takes on only a finite number of values in any finite interval. For example, when examining a policyholder's choice of deductibles, it may be that values of 0, 250, 500, and 1000 are the only possible outcomes. Like an ordinal variable, these represent distinct categories that are ordered. Unlike an ordinal variable, the numerical difference between levels takes on economic meaning. A special type of discrete variable is a count variable, one with values on the nonnegative integers. For example, we will be particularly interested in the number of claims arising from a policy during a given period. Another interesting variation is an interval variable, one that gives a range of possible outcomes.

Earlier, in the Section 1.3 case study, we encountered several examples of quantitative variables. These included the deductible (in logarithmic dollars), total building and content coverage (in logarithmic dollars), claim severity and claim frequency.

### Loss Data

This introduction to data analytics is motivated by features of **loss data** that arise from, or are related to, obligations in insurance contracts. Loss data rarely arise from a bell-shaped normal distribution that has motivated the development of much of classical statistics. As a consequence, the treatment of data analytics in this text differs from that typically encountered in other introductions to data analytics.

What features of loss data warrant special treatment?

- We have already seen in the Section 1.3 case study that we will be concerned with the frequency of losses, a type of count variable.
- Further, when a loss occurs, the interest is in the amount of the claim, a quantitative variable. This claim severity is commonly modeled using skewed and long-tailed distributions so that extremely large outcomes are associated with relatively large probabilities. Typically, the normal distribution is a poor choice for a loss distribution.
- When a loss does occur, often the analyst only observes a value that is modified by insurance contractual features such as deductibles, upper limits, and co-insurance parameters.
- Loss data are frequently a *combination of discrete and continuous components*. For example, when we analyze the insured loss of a policyholder, we will encounter a discrete outcome at zero, representing no insured loss, and a continuous amount for positive outcomes, representing the amount of the insured loss.

#### 2.3.2 Exploratory versus Confirmatory

There are two phases of data analysis: exploratory data analysis (EDA) and confirmatory data analysis (CDA). Table 2.3 summarizes some differences between EDA and CDA. EDA is usually applied to observational data with the goal of looking for patterns and formulating hypotheses. In contrast, CDA is often applied to experimental data (i.e., data obtained by means of a formal design of experiments) with the goal of quantifying the extent to which discrepancies between the model and the data could be expected to occur by chance.

**Table 2.3. Comparison of Exploratory Data Analysis and Confirmatory Data Analysis**

	<b>EDA</b>	<b>CDA</b>
Data	Observational data	Experimental data
Goal	Pattern recognition, formulate hypotheses	Hypothesis testing, estimation, prediction
Techniques	Descriptive statistics, visualization, clustering	Traditional statistical tools of inference, significance, and confidence

As we have seen in the Section 1.3 case study, the techniques for single variable EDA include descriptive statistics (e.g., mean, median, standard deviation, quantiles) and summaries of distributions such as through histograms. In contrast, the techniques for CDA include the traditional statistical tools of inference, significance, and confidence.

### 2.3.3 Model Construction

As we learned in Section 2.1.2, models may have a stochastic basis from the statistical modeling paradigm or may simply be the result of an algorithm. When constructing a model, it is helpful to think about how it is parameterized and to identify the purpose of constructing the model.

#### Parametric versus Nonparametric

Data analysis models can be parametric or nonparametric. Parametric models are representations that are known up to a few terms known as *parameters*. These may be representations of a stochastic distribution or simply an algorithm used to predict data outcomes. Typically, data are used to determine the parameters and in this way calibrate the model. In contrast, nonparametric methods make no such assumption of a known functional form. For example, Section 4.4.1 will introduce nonparametric methods that do not assume distributions for the data and therefore are also called *distribution-free* methods.

Because a functional form is known with a parametric model, this approach works well when data size is relatively limited. This reasoning extends to the situation when one is considering many variables simultaneously so that the so-called “curse of dimensionality” effectively limits the sample size. For example if you are trying to determine the expected cost of automobile losses, you are likely to consider a driver’s age, gender, driving location, type of vehicle, and dozens of other variables. Approaches that use some parametric relationships

among these variables are common because a purely non-parametric approach would require data sets too large to be useful in practice.

Nonparametric methods are very valuable particularly at the exploratory stages of an analysis where one tries to understand the distribution of each variable. Because nonparametric methods make fewer assumptions, they can be more flexible, more robust, and more applicable to non-quantitative data. However, a drawback of nonparametric methods is that it is more difficult to extrapolate findings outside of the observed domain of the data, a key consideration in *predictive modeling*.

### Explanation versus Prediction

There are two goals in data analysis: explanation and prediction. In some scientific areas such as economics, psychology, and environmental science, the focus of data analysis is to explain the causal relationships between the input variables and the response variable. In other scientific areas such as natural language processing, bioinformatics, and actuarial science, the focus of data analysis is to predict what the responses are going to be given the input variables.

[Shmueli \(2010\)](#) discussed in detail the distinction between explanatory modeling and predictive modeling. Explanatory modeling is commonly used for theory building and testing and is typically done as follows:

- State the prevailing theory.
- State causal hypotheses, which are given in terms of theoretical constructs rather than measurable variables. A causal diagram is usually included to illustrate the hypothesized causal relationship between the theoretical constructs.
- Operationalize constructs. In this step, previous literature and theoretical justification are used to build a bridge between theoretical constructs and observable measurements.
- Collect data and build models alongside the statistical hypotheses, which are operationalized from the research hypotheses.
- Reach research conclusions and recommend policy. The statistical conclusions are converted into research conclusions or policy recommendations.

In contrast, predictive modeling is the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations. Predictions include point predictions, interval predictions, regions, distributions, and rankings of new observations. A predictive model can be any method that produces predictions.

### 2.3.4 Model Selection

Although hypothesis testing is one approach to model selection that is viable in many fields, it does have its drawbacks. For example, the asymmetry between the null and alternative hypotheses raises issues; hypothesis testing is biased towards a null hypothesis unless there is strong evidence to the contrary.

For modeling insurance activities, it is typically preferable to estimate the predictive power of various models and select a model with the best predictive power. The motivation for this is that we want good model selection methods achieve a balance between goodness of fit and parsimony. This is a trade-off because on the one hand, better fits to the data can be achieved by adding more parameters, making the model more complex and less parsimonious. On the other hand, models with fewer parameters (parsimonious) are attractive because of their simplicity and interpretability; they are also less subject to estimation variability and so can yield more accurate predictions, [Ruppert et al. \(2003\)](#).

One way of measuring this balance is through information criteria such as Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These measures each contain a component that summarizes how well the model fits the data, a goodness of fit piece, plus a component to penalize the complexity of the model.

Although attractive due to their simplicity, there are drawbacks to these measures. In particular, both rely on knowledge of the underlying distribution of the outcomes (or at least good estimates). A more robust approach is to split a data set in a portion that can be used to calibrate a model, the *training* portion, and another portion used to quantify the predictive power of the model, the *test* portion. It is more robust in the sense that it does not rely on any distributional assumptions and can be used to validate general models.

The data splitting approach is attractive because it directly aligns with the concept of assessing predictive power and can be used in general, and complex, situations. However, it does introduce additional variability into the process by introducing extra randomness of the uncertainty of which observations fall into the training and testing portions. To mitigate this problem, it is common to use an approach known as *cross-validation*. To illustrate, suppose that one randomly partitions a dataset into five subsets of roughly equivalent sizes



Then, based on the first, third, fourth, and fifth subsets, estimate a model, use

this fitted model to predict outcomes in the second, and compare the predictions to the held-out values in the test portion. Repeat this process by selecting each subset as the test portion, with the others being used for training, and take an average over the comparison which results in a cross-validation statistic. Cross-validation is used widely in modeling insurance activities and is described in more detail in Chapter 5.

**Example 2.3.1. Under- and Over-Fitting.** Suppose that we have a set of claims that potentially varies by a single categorical variable with six levels. For example, in the Section 1.3 case study there are six entity types. If each level is truly distinct, then in classical statistics one uses the level average to make predictions for future claims. Another option is to ignore information in the categorical variable and use the overall average to make predictions; this is known as a “community-rating” approach.

For illustrative purposes, we assume that two of the six levels are the same and are different from the others. For example, the Table 1.6 summary statistics suggest that Schools and the Miscellaneous levels can be viewed similarly yet warrant a higher predicted claims amount than the other four levels. For illustrative purposes, we generated 100 claims that follow this pattern (using simulation techniques that will be described in Chapter 8).

Results are summarized in Table 2.4 for three fitted models. These are the “Community Rating” corresponding to using the overall mean, the “Two Levels” corresponding to using two averages, and the “Six Levels” corresponding to using an average for each level of the categorical variable. The data set of size 100 was randomly split into five folds; for each fold, the other folds were used to train/estimate the model and then that fold was used to assess predictions. The first five rows of Table 2.4 give the results of the root mean square error for each fold. The sixth row provides the average over the five folds and the last row gives a similar result for another goodness of fit statistic, the *AIC*. This approach is known as “cross-validation” that will be described in greater detail in Chapters 6 and 8.

Table 2.4 shows that in each case the “Two Level” model has the lowest root mean square error and *AIC*, indicating that it is the preferred model. The overfit model with six levels came in second and the underfit model, community rating, was a distant third. This analysis demonstrates techniques for selecting the appropriate model. Unlike analysis of real data, in this demonstration we enjoyed the additional luxury of knowing that we got things correct because we in fact generated the data - an approach that analysts often use to develop analytic procedures prior to utilizing the procedures on real data.

---

TABLE 2.4: Under- and Over-Fitting of Models

	Community Rating	Two Levels	Six Levels
Rmse - Fold 1	1.318	1.192	1.239
Rmse - Fold 2	1.034	0.972	1.023
Rmse - Fold 3	0.816	0.660	0.759
Rmse - Fold 4	0.807	0.796	0.824
Rmse - Fold 5	0.886	0.539	0.671
Rmse - Average	0.972	0.832	0.903
AIC - Average	227.171	206.769	211.333

## 2.4 Analytics with Many Variables

In this section, you learn how to describe analytics based on many variables in terms of

- supervised and unsupervised learning,
- types of algorithmic models, including linear, ridge, and lasso regressions, as well as regularization, and
- types of data models, including Poisson regressions and generalized linear models.

Just as with a single variable in Section 2.3, with many variables analysts follow the same structure of identifying variables, exploring data, constructing and selecting models. However, the potential applications become much richer when considering many variables. With many potential applications, it is natural that techniques for data analysis have developed in different but overlapping fields; these fields include statistics, machine learning, pattern recognition, and data mining.

- Statistics is a field that addresses reliable ways of gathering data and making inferences.
- The term machine learning was coined by Samuel in 1959 ([Samuel, 1959](#)). Originally, machine learning referred to the field of study where computers have the ability to learn without being explicitly programmed. Nowadays, machine learning has evolved to a broad field of study where computational methods use experience (i.e., the past information available for analysis) to improve performance or to make accurate predictions.

- Originating in engineering, pattern recognition is a field that is closely related to machine learning, which grew out of computer science. In fact, pattern recognition and machine learning can be considered to be two facets of the same field ([Bishop, 2007](#)).
- Data mining is a field that concerns collecting, cleaning, processing, analyzing, and gaining useful insights from data ([Aggarwal, 2015](#)).

### 2.4.1 Supervised and Unsupervised Learning

With multiple variables, the essential tasks of identifying variable types, exploring data, and selecting models are similar in principle to that described for single variables in Section 2.3. When exploring data in multiple dimensions, additional considerations such as clustering like observations and reducing the dimension arise. As these considerations will not arise in the applications in this book, we provide only a brief introduction in Technical Supplement Section 2.6.1.

The construction of models differs dramatically when comparing single to multiple variable modeling. With many variables, we have the opportunity to think about some of them as “inputs” and others “outputs” of a system. Models based on input and output variables are known as supervised learning methods or as regression methods. [Table 2.5](#) gives a list of common names for different types of variables ([Frees, 2009](#)). When the target variable is a categorical variable, supervised learning methods are called classification methods.

Table 2.5. Common Names of Different Variables

Target Variable	Explanatory Variable
Dependent variable	Independent variable
Response	Treatment
Output	Input
Endogenous variable	Exogenous variable
Predicted variable	Predictor variable
Regressand	Regressor

Methods for data analysis can be divided into two types ([Abbott, 2014](#); [James et al., 2013](#)): supervised learning methods and unsupervised learning methods. Unsupervised learning methods work where our data are treated the same and there is no artificial divide between “inputs” and “outputs.” As a result, unsupervised learning methods are particularly useful at the exploratory stage of an analysis.

### 2.4.2 Algorithmic Modeling

Early data analysis traced the movements of orbits of bodies about the sun using astronomical observations in the 1750's by Boscovich and was continued in the early 1800's by Legendre and Gauss (the latter two in connection with their development of least squares). This work was done using algorithmic *fitting* approaches (such as least squares) without regard to distributions of random variables.

The idea underpinning algorithmic fitting is easy to interpret. One variable,  $Y$ , is determined to be a target variable. Other variables,  $X_1, X_2, \dots, X_p$ , are used to understand or explain the target  $Y$ . The goal is to determine an appropriate function  $f(\cdot)$  so that  $f(X_1, X_2, \dots, X_k)$  is a useful predictor of  $Y$ .

**Linear Regression.** To illustrate, consider the classic linear regression context. In this case, we have  $n$  observations of a target and explanatory variables, with the  $i$ th observation denoted as  $(x_{i1}, \dots, x_{ik}, y_i) = (\mathbf{x}_i, y_i)$ . One would like to determine a single function  $f$  so that  $f(\mathbf{x}_i)$  is a reasonable approximation for  $y_i$ , for each  $i$ . For the linear regression, one restricts considerations to functions of the form

$$f(x_{i1}, \dots, x_{ik}) = \beta_1 x_{i1} + \dots + \beta_k x_{ik} = \mathbf{x}'_i \boldsymbol{\beta}.$$

Here,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$  is a vector of *regression coefficients*. This function is *linear* in the explanatory variables that gives rise to the name linear regression.

The *ordinary least squares (OLS)* estimates are the solution of the following minimization problem,

$$\text{minimize}_{\boldsymbol{\beta}} \quad \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2.$$

The *OLS* estimates are historically prominent in part because of their ease of computation and interpretation. Naturally, a squared difference such as  $(y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$  is not the only way to measure the deviation between a target  $y_i$  and an estimate  $\mathbf{x}'_i \boldsymbol{\beta}$ . In general, analysts use the term *loss function*  $l(y_i, \mathbf{x}'_i \boldsymbol{\beta})$  to measure this deviation; as an alternative, it is not uncommon to use an absolute deviation.

**Algorithmic Modeling Culture.** As introduced in Section 2.1.2, a culture has developed across widespread communities that emphasizes algorithmic fitting particularly in complex problems such as voice, image, and handwriting recognition. Algorithmic methods are especially useful when the goal is prediction, as noted in Section 2.3.3. Many of these algorithms take an approach similar to linear regression. As examples, other widely used algorithmic fitting methods include ridge and lasso regression, as well as regularization methods.

**Ridge Regression.** One limitation of *OLS* is that it tends to overfit, particularly when the number of regression coefficients  $k$  becomes large. In fact, with  $k = n$  one gets an exact match between the targets  $y_i$  and the predictor function. A modification introduced in 1970 by Hoerl and Kennard (1970) is known as *ridge regression* where one determines regression coefficients  $\beta$  as in equation (2.4.2) although subject to the constraint that  $\sum_{j=1}^p |\beta_j|^2 \leq c_{ridge}$ , where  $c_{ridge}$  is an appropriately chosen constant. Naturally, if  $c_{ridge}$  is very large, then the constraint has no effect and the ridge estimates equal the *OLS* solution. However, as  $c_{ridge}$  becomes small, it reduces the size of the regression coefficients. In this sense, the ridge regression estimator is said to be “shrunk towards zero.”

Adding the constraint on the size of the coefficients can mean smaller and more stable coefficients when compared to *OLS*. As such, ridge regression is particularly useful when dealing with high-dimensional datasets, where the number of predictors is very large compared to the number of observations. In the actuarial applications, we might have a portfolio of only a few thousand risks that we wish to model. With ridge regression, we can utilize millions of variables as potential inputs to develop predictive models.

**Lasso Regression.** Similar to ridge regression, one can determine regression coefficients  $\beta$  as in equation (2.4.2) although subject to the constraint that  $\sum_{j=1}^p |\beta_j| \leq c_{lasso}$ , where  $c_{lasso}$  is an appropriately chosen constant. This procedure is known as *lasso regression*. Here, one uses absolute values in the constraint function (although still squared errors for the loss function).

The lasso overcomes an important limitation of ridge regression. With ridge regression, we might reduce the size of the constant  $c_{ridge}$  that forces the regression coefficient to become small but does not ensure that they become zero. In contrast, the lasso ensures that trivial regression coefficients become zero. In the linear regression approximation, a zero regression coefficient means that the variable drops from the function approximation, thus reducing model complexity.

**Regularization.** Both the ridge and lasso regressions are constrained minimization problems. It is not too hard to show that they can be written as

$$\text{minimize}_{\beta} \left( \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}'_i \beta)^2 + LM \sum_{j=1}^p |\beta_j|^s \right),$$

where  $s = 2$  is for ridge regression and  $s = 1$  is for lasso regression. We can interpret the first part inside the minimization operation as the goodness of fit and the second part as a penalty for size of the regression coefficients. As we have discussed, reducing the coefficients can mean reducing modeling complexity. In this sense, this expression demonstrates a balance between goodness

of fit and model complexity, controlled by the parameter  $LM$  (In this case, because it is a constrained optimization problem, the parameter is a Lagrange multiplier.). The choice of  $LM = 0$  reduces to the *OLS* estimator that focuses on goodness of fit. As  $LM$  becomes large, the focus moves away from the data (and hence goodness of fit). This is an example of a *regularization* method in data analytics, where one expresses a prior belief concerning the smoothness of functions used for our predictions.

### 2.4.3 Data Modeling

One way to motivate an algorithmic development is through the use of a data model introduced in Section 2.1.2. Here, we can also think of this as a “probability” or “likelihood” based model, in that our main goal is to understand the target ( $Y$ ) distribution, typically in terms of the explanatory variables. Thus, data models are particularly useful for the goal of explanation previously discussed in Section 2.3.3.

Data models were initially developed in the early twentieth century through the work of R.A. Fisher and E.P. George Box (among many, many others) whose work focused on data as the result of experiments with a small number of outcomes and even fewer explanatory (control) variables.

**Linear Regression.** The (algorithmic) linear regression with *OLS* estimates can be motivated using a probabilistic framework, as follows. We can think of the target variable  $y_i$  as having a normal distribution with unknown variance and a mean equal to  $\mathbf{x}'_i\beta$ , a linear combination of the explanatory variables. Assuming independence among observations, it can be shown that the maximum likelihood estimates are equivalent to the *OLS* estimates determine in equation (2.4.2).

Maximum likelihood estimation is used extensively in this text, *you can get a quick overview in Chapter 18 Appendix C*. For additional background on *OLS* and maximum likelihood in the linear regression case see, for example, [Frees \(2009\)](#) for more details.

**Poisson Regression.** In the case where the target variable  $Y$  represents a count (such as the number of insurance losses), then it is common to use a Poisson distribution to represent the likelihood of potential outcomes. The Poisson has only one parameter, the mean, and if explanatory variables are available, then one can take the mean to equal  $\exp(\mathbf{x}'_i\beta)$ . One motivation for using the exponential ( $\exp(\cdot)$ ) function is that it ensures that estimated means are non-negative (a necessary condition for the Poisson distribution). When maximum likelihood is used to estimate the regression coefficients, then this is known as *Poisson regression*.

**Generalized Linear Model.** The generalized linear model (*GLM*) consists of a wide family of regression models that include linear and Poisson regression models as special cases. In a *GLM*, the mean of the target variable is assumed to be a function of a linear combination of the explanatory variables. As with a Poisson regression, the mean can vary by observations by allowing some parameters to change yet the regression parameters  $\beta$  are assumed to be constant.

In a *GLM*, the target variable is assumed to follow a distribution from the *linear exponential family*, a collection of distributions that includes the normal, Poisson, Bernoulli, Weibull, and others. Thus, a *GLM* is one way of developing a broader class that includes linear and Poisson regression. Using a Bernoulli distribution, it also includes zero-one target variables resulting in what is known as *logistic regression*. Thus, the *GLM* provides a unifying framework to handle different types of target variables, including discrete and continuous variables. Extensions to other distributions that are not part of linear exponential family, such as a Pareto distribution, are also possible. But, *GLMs* have historically been found useful because their form permits efficient calculation of estimators (through what is known as *iterative reweighted least squares*). For more information about *GLMs*, readers are referred to [De Jong and Heller \(2008\)](#) and [Frees \(2009\)](#).

---

## 2.5 Data

---

In this section, you learn how to describe data considerations in terms of

- data types,
  - data structure and storage,
  - data cleaning,
  - big data issues, and
  - ethical issues.
- 

Data constitute the backbone of “data analytics.” Without data containing useful information, no level of sophisticated analytic techniques can provide useful guidance for making good decisions.

The prior sections of this chapter provide the foundations of data considerations needed for the rest of this book. However, for readers who wish to

specialize in data analytics, the following subsections provide a useful starting point for further study.

### 2.5.1 Data Types

In terms of how data are collected, data can be divided into two types ([Hox and Boeije, 2005](#)): primary and secondary data. Primary data are the original data that are collected for a specific research problem. Secondary data are data originally collected for a different purpose and reused for another research problem. A major advantage of using primary data is that the theoretical constructs, the research design, and the data collection strategy can be tailored to the underlying research question to ensure that data collected help to solve the problem. A disadvantage of using primary data is that data collection can be costly and time consuming. Using secondary data has the advantage of lower cost and faster access to relevant information. However, using secondary data may not be optimal for the research question under consideration.

In terms of the degree of organization, data can be also divided into two types: structured data and unstructured data. Structured data have a predictable and regularly occurring format. In contrast, unstructured data lack any regularly occurring format and have no structure that is recognizable to a computer. Structured data consist of records, attributes, keys, and indices and are typically managed by a database management system such as IBM DB2, Oracle, MySQL, and Microsoft SQL Server. As a result, most units of structured data can be located quickly and easily. Unstructured data have many different forms and variations. One common form of unstructured data is text. Accessing unstructured data can be awkward. To find a given unit of data in a long text, for example, a sequential search is usually performed.

### 2.5.2 Data Structures and Storage

As mentioned in the previous subsection, there are structured data as well as unstructured data. Structured data are highly organized data and usually have the following tabular format:

	$V_1$	$V_2$	$\dots$	$V_d$
$\mathbf{x}_1$	$x_{11}$	$x_{12}$	$\dots$	$x_{1d}$
$\mathbf{x}_2$	$x_{21}$	$x_{22}$	$\dots$	$x_{2d}$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$
$\mathbf{x}_n$	$x_{n1}$	$x_{n2}$	$\dots$	$x_{nd}$

In other words, structured data can be organized into a table consisting of rows and columns. Typically, each row represents a record and each column

represents an attribute. A table can be decomposed into several tables that can be stored in a relational database such as the Microsoft SQL Server. The SQL (Structured Query Language) can be used to access and modify the data easily and efficiently.

Unstructured data do not follow a regular format. Examples of unstructured data include documents, videos, and audio files. Most of the data we encounter are unstructured data. In fact, the term “big data” was coined to reflect this fact. Traditional relational databases cannot meet the challenges on the varieties and scales brought by massive unstructured data nowadays. NoSQL databases have been used to store massive unstructured data.

There are three main NoSQL databases ([Chen et al., 2014](#)): key-value databases, column-oriented databases, and document-oriented databases. Key-value databases use a simple data model and store data according to key values. Modern key-value databases have higher expandability and smaller query response times than relational databases. Examples of key-value databases include Dynamo used by Amazon and Voldemort used by LinkedIn. Column-oriented databases store and process data according to columns rather than rows. The columns and rows are segmented in multiple nodes to achieve expandability. Examples of column-oriented databases include BigTable developed by Google and Cassandra developed by Facebook. Document databases are designed to support more complex data forms than those stored in key-value databases. Examples of document databases include MongoDB, SimpleDB, and CouchDB. MongoDB is an open-source document-oriented database that stores documents as binary objects. SimpleDB is a distributed NoSQL database used by Amazon. CouchDB is another open-source document-oriented database.

### 2.5.3 Data Cleaning

Raw data usually need to be cleaned before useful analysis can be conducted. In particular, the following areas need attention when preparing data for analysis ([Janert, 2010](#)):

- **Missing values.** It is common to have missing values in raw data. Depending on the situation, we can discard the record, discard the variable, or impute the missing values.
- **Outliers.** Raw data may contain unusual data points such as outliers. We need to handle outliers carefully. We cannot just remove outliers without knowing the reason for their existence. Although sometimes outliers can be simple mistakes such as those caused by clerical errors, sometimes their unusual behavior can point to precisely the type of effect that we are looking for.

- **Junk.** Raw data may contain garbage, or junk, such as nonprintable characters. When it happens, junk is typically rare and not easily noticed. However, junk can cause serious problems in downstream applications.
- **Format.** Raw data may be formatted in a way that is inconvenient for subsequent analysis. For example, components of a record may be split into multiple lines in a text file. In such cases, lines corresponding to a single record should be merged before loading to a data analysis software such as R.
- **Duplicate records.** Raw data may contain duplicate records. Duplicate records should be recognized and removed. This task may not be trivial depending on what you consider “duplicate.”
- **Merging datasets.** Raw data may come from different sources. In such cases, we need to merge data from different sources to ensure compatibility.

For more information about how to handle data in R, readers are referred to [Forte \(2015\)](#) and [Buttrey and Whitaker \(2017\)](#).

#### 2.5.4 Big Data Analysis

Unlike traditional data analysis, big data analysis employs additional methods and tools that can extract information rapidly from massive data. In particular, big data analysis uses the following processing methods ([Chen et al., 2014](#)):

- A **bloom filter** is a space-efficient probabilistic data structure that is used to determine whether an element belongs to a set. It has the advantages of high space efficiency and high query speed. A drawback of using bloom filter is that there is a certain nonrecognition rate.
- **Hashing** is a method that transforms data into fixed-length numerical values through a hash function. It has the advantages of rapid reading and writing. However, sound hash functions are difficult to find.
- **Indexing** refers to a process of partitioning data in order to speed up reading. Hashing is a special case of indexing.
- A **trie**, also called digital tree, is a method to improve query efficiency by using common prefixes of character strings to reduce comparisons among character strings.
- **Parallel computing** uses multiple computing resources to complete a computation task. Parallel computing tools include Message Passing Interface (MPI), MapReduce, and Dryad.

Big data analysis can be conducted in the following levels ([Chen et al., 2014](#)): memory-level, business intelligence (BI) level, and massive level. Memory-level analysis is conducted when data can be loaded to the memory of a cluster of computers. Current hardware can handle hundreds of gigabytes (GB) of data in memory. BI level analysis can be conducted when data surpass the memory

level. It is common for BI level analysis products to support data over terabytes (TB). Massive level analysis is conducted when data surpass the capabilities of products for BI level analysis. Usually Hadoop and MapReduce are used in massive level analysis.

### 2.5.5 Ethical Issues

Analysts may face ethical issues and dilemmas during the data analysis process. In some fields, ethical issues and dilemmas include participant consent, benefits, risk, confidentiality, and data ownership (Miles et al., 2014). For example, regarding privacy and confidentiality, one might confront the following questions: How do we make sure that the information is kept confidentially? How do we verify where raw data and analysis results are stored? How will we have access to them? These questions should be addressed and documented in explicit confidentiality agreements.

Within the insurance sector, discrimination, privacy, and confidentiality are major concerns. Discrimination in insurance is particularly difficult because the entire industry is based on “discriminating,” or classifying, insureds into homogeneous categories for the purposes of risk sharing. Many variables that insurers use are seemingly innocuous (e.g., blindness for auto insurance), yet others can be viewed as “wrong” (e.g., religious affiliation), “unfair” (e.g., onset of cancer for health insurance), “sensitive” (e.g., marital status), or “mysterious” (e.g., Artificial Intelligence produced). Regulators and policymakers decide whether it is not permitted to use a variable for classification. In part because they depend on differing attitudes, perspectives can vary dramatically across jurisdictions. For example, gender-based pricing of auto insurance is permitted in all but a handful of U.S. states (the exceptions being Hawaii, Massachusetts, Montana, North Carolina, Pennsylvania, and, as of 2019, California) yet not permitted within the European Union. Moreover, for personal lines such as auto and homeowners, availability of big data may also lead to issues regarding *proxy discrimination*. Proxy discrimination occurs when a surrogate, or proxy, is used in place of a prohibited trait such as race or gender, see, for example, [Frees and Huang \(2021\)](#).

---

## 2.6 Further Resources and Contributors

### Contributors

- **Guojun Gan**, University of Connecticut, was the principal author of the initial version of this chapter.

- Chapter reviewers include: Runhuan Feng, Himchan Jeong, Lei Hua, Min Ji, and Toby White.
- **Hirokazu (Iwahiro) Iwasawa** and **Edward (Jed) Frees**, University of Wisconsin-Madison and Australian National University, are the authors of the second edition of this chapter. Email: [iwahiro@bb.mbn.or.jp](mailto:iwahiro@bb.mbn.or.jp) and/or [jfrees@bus.wisc.edu](mailto:jfrees@bus.wisc.edu) for chapter comments and suggested improvements.

### Further Readings and References

- [Stigler \(1986\)](#) gives a definitive account of the early contributions of Boscovich, Legendre and Gauss.
- [Breiman \(2001\)](#) compares the data modeling and the algorithmic modeling cultures.
- [Good \(1983\)](#) compares the two phases of data analysis, exploratory data analysis (EDA) and confirmatory data analysis (CDA)
- See, for example, [Breiman \(2001\)](#) and [Shmueli \(2010\)](#), for more discussions of the two goals in data analysis: explanation and prediction.
- Comparisons of structured data and unstructured data can be found in [Inmon and Linstedt \(2014\)](#), [O'Leary \(2013\)](#), [Hashem et al. \(2015\)](#), [Abdullah and Ahmad \(2013\)](#), and [Pries and Dunnigan \(2015\)](#), among others.

#### 2.6.1 Technical Supplement: Multivariate Exploratory Analysis

##### Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that transforms a dataset described by possibly correlated variables into a dataset described by linearly uncorrelated variables, which are called principal components and are ordered according to their variances. PCA is a technique for dimension reduction. If the original variables are highly correlated, then the first few principal components can account for most of the variation of the original data.

The principal components of the variables are related to the eigenvalues and eigenvectors of the covariance matrix of the variables. For  $i = 1, 2, \dots, d$ , let  $(\lambda_i, \mathbf{e}_i)$  be the  $i$ th eigenvalue-eigenvector pair of the covariance matrix  $\Sigma$  of  $d$  variables  $X_1, X_2, \dots, X_d$  such that  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$  and the eigenvectors are normalized. Then the  $i$ th principal component is given by

$$Z_i = \mathbf{e}'_i \mathbf{X} = \sum_{j=1}^d e_{ij} X_j,$$

where  $\mathbf{X} = (X_1, X_2, \dots, X_d)'$ . It can be shown that  $\text{Var}(Z_i) = \lambda_i$ . As a result, the proportion of variance explained by the  $i$ th principal component is

calculated as

$$\frac{\text{Var}(Z_i)}{\sum_{j=1}^d \text{Var}(Z_j)} = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \dots + \lambda_d}.$$

For more information about PCA, readers are referred to [Mirkin \(2011\)](#).

### Cluster Analysis

Cluster analysis (aka data clustering) refers to the process of dividing a dataset into homogeneous groups or clusters such that points in the same cluster are similar and points from different clusters are quite distinct ([Gan et al., 2007](#); [Gan, 2011](#)). Data clustering is one of the most popular tools for exploratory data analysis and has found its applications in many scientific areas.

During the past several decades, many clustering algorithms have been proposed. Among these clustering algorithms, the  $k$ -means algorithm is perhaps the most well-known algorithm due to its simplicity. To describe the  $k$ -means algorithm, let  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  be a dataset containing  $n$  points, each of which is described by  $d$  numerical features. Given a desired number of clusters  $k$ , the  $k$ -means algorithm aims at minimizing the following objective function:

$$P(U, Z) = \sum_{l=1}^k \sum_{i=1}^n u_{il} \|\mathbf{x}_i - \mathbf{z}_l\|^2,$$

where  $U = (u_{il})_{n \times k}$  is an  $n \times k$  partition matrix,  $Z = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k\}$  is a set of cluster centers, and  $\|\cdot\|$  is the  $L^2$  norm or Euclidean distance. The partition matrix  $U$  satisfies the following conditions:

$$u_{il} \in \{0, 1\}, \quad i = 1, 2, \dots, n, \quad l = 1, 2, \dots, k,$$

$$\sum_{l=1}^k u_{il} = 1, \quad i = 1, 2, \dots, n.$$

The  $k$ -means algorithm employs an iterative procedure to minimize the objective function. It repeatedly updates the partition matrix  $U$  and the cluster centers  $Z$  alternately until some stop criterion is met. For more information about  $k$ -means, readers are referred to [Gan et al. \(2007\)](#) and [Mirkin \(2011\)](#).

#### 2.6.2 Tree-based Models

Decision trees, also known as tree-based models, involve dividing the predictor space (i.e., the space formed by independent variables) into a number of simple regions and using the mean or the mode of the region for prediction ([Breiman et al., 1984](#)). There are two types of tree-based models: classification trees and regression trees. When the dependent variable is categorical, the resulting

tree models are called classification trees. When the dependent variable is continuous, the resulting tree models are called regression trees.

The process of building classification trees is similar to that of building regression trees. Here we only briefly describe how to build a regression tree. To do that, the predictor space is divided into non-overlapping regions such that the following objective function

$$f(R_1, R_2, \dots, R_J) = \sum_{j=1}^J \sum_{i=1}^n I_{R_j}(\mathbf{x}_i)(y_i - \mu_j)^2$$

is minimized, where  $I$  is an indicator function,  $R_j$  denotes the set of indices of the observations that belong to the  $j$ th box,  $\mu_j$  is the mean response of the observations in the  $j$ th box,  $\mathbf{x}_i$  is the vector of predictor values for the  $i$ th observation, and  $y_i$  is the response value for the  $i$ th observation.

In terms of predictive accuracy, decision trees generally do not perform to the level of other regression and classification models. However, tree-based models may outperform linear models when the relationship between the response and the predictors is nonlinear. For more information about decision trees, readers are referred to [Breiman et al. \(1984\)](#) and [Mitchell \(1997\)](#).

### 2.6.3 Technical Supplement: Some R Functions

R is an open-source software for statistical computing and graphics. The R software can be downloaded from the R project website at <https://www.r-project.org/>. In this section, we give some R function for data analysis, especially the data analysis tasks mentioned in previous sections.

Table 2.6. Some R Functions for Data Analysis

Data Analysis Task	R Package	R Function
Descriptive Statistics	base	summary
Principal Component Analysis	stats	prcomp
Data Clustering	stats	kmeans, hclust
Fitting Distributions	MASS	fitdistr
Linear Regression Models	stats	lm
Generalized Linear Models	stats	glm
Regression Trees	rpart	rpart
Survival Analysis	survival	survfit

**Table 2.6** lists a few R functions for different data analysis tasks. Readers can go to the R documentation to learn how to use these functions. There are also other R packages that do similar things. However, the functions listed in this table provide good starting points for readers to conduct data analysis in R. For analyzing large datasets in R in an efficient way, readers are referred to [Daroczi \(2015\)](#).

---

This work is licensed under a Creative Commons Attribution 4.0 International License.

# 3

---

## *Frequency Modeling*

---

*Chapter Preview.* A primary focus for insurers is estimating the magnitude of aggregate claims it must bear under its insurance contracts. Aggregate claims are affected by both the frequency and the severity of the insured event. Decomposing aggregate claims into these two components, each of which warrant significant attention, is essential for analysis and pricing. This chapter discusses frequency distributions, summary measures, and parameter estimation techniques.

In Section 3.1, we present terminology and discuss reasons why we study frequency and severity separately. The foundations of frequency distributions and measures are presented in Section 3.2 along with three principal distributions: the binomial, the Poisson, and the negative binomial. These three distributions are members of what is known as the  $(a, b, 0)$  class of distributions, a distinguishing, identifying feature which allows for efficient calculation of probabilities, further discussed in Section 3.3. When fitting a dataset with a distribution, parameter values need to be estimated and in Section 3.4, the procedure for maximum likelihood estimation is explained.

For insurance datasets, the observation at zero denotes no occurrence of a particular event; this often deserves additional attention. As explained further in Section 3.5, for some datasets it may be impossible to have zero of the studied event or zero events may follow a different model than other event counts. In either case, direct fitting of typical count models could lead to improper estimates. Zero truncation or modification techniques allow for more appropriate distribution fit.

Noting that our insurance portfolio could consist of different sub-groups, each with its own set of individual characteristics, Section 3.6 introduces mixture distributions and methodology to allow for this heterogeneity within a portfolio. In Section 3.7 an example is given that demonstrates how standard frequency distributions can often provide a good fit to real data. Exercises are presented in Section 3.8 and Section 3.9.1 concludes the chapter with R Code for plots depicted in Section 3.4.

### 3.1 Frequency Distributions

---

In this section, you learn how to summarize the importance of frequency modeling in terms of

- contractual,
  - behavioral,
  - database, and
  - regulatory/administrative motivations.
- 

#### 3.1.1 How Frequency Augments Severity Information

##### Basic Terminology

In this chapter, **loss**, also referred to as ground-up loss, denotes the amount of financial loss suffered by the insured. We use **claim** to denote the indemnification upon the occurrence of an insured event, thus the amount paid by the insurer. While some texts use **loss** and **claim** interchangeably, we wish to make a distinction here to recognize how insurance contractual provisions, such as deductibles and limits, affect the size of the claim stemming from a loss. Frequency represents how often an insured event occurs, typically within a policy contract. Here, we focus on count random variables that represent the number of claims, that is, how frequently an event occurs. Severity denotes the amount, or size, of each payment for an insured event. In Chapter 7, the aggregate model, which combines frequency models with severity models, is examined.

##### The Importance of Frequency

Recall from Section 1.2 that setting the price of an insurance good can be a complex problem. In manufacturing, the cost of a good is (relatively) known. In other financial service areas, market prices are available. In insurance, we can generalize the price setting as follows. Start with an expected cost, then add “margins” to account for the product’s riskiness, expenses incurred in servicing the product, and a profit/surplus allowance for the insurer.

The expected cost for insurance can be determined as the expected number of claims times the amount per claim, that is, expected value of *frequency times severity*. The focus on claim count allows the insurer to consider those factors

which directly affect the occurrence of a loss, thereby potentially generating a claim.

### Why Examine Frequency Information?

Insurers and other stakeholders, including governmental organizations, have various motivations for gathering and maintaining frequency datasets.

- **Contractual.** In insurance contracts, it is common for particular deductibles and policy limits to be listed and invoked for each occurrence of an insured event. Correspondingly, the claim count data generated would indicate the number of claims which meet these criteria, offering a unique claim frequency measure. Extending this, models of total insured losses would need to account for deductibles and policy limits for each insured event.
- **Behavioral.** In considering factors that influence loss frequency, the risk-taking and risk-reducing behavior of individuals and companies should be considered. Explanatory (rating) variables can have different effects on models of how often an event occurs in contrast to the size of the event.
  - In healthcare, the decision to utilize healthcare by individuals, and minimize such healthcare utilization through preventive care and wellness measures, is related primarily to his or her personal characteristics. The cost per user is determined by the patient's medical condition, potential treatment measures, and decisions made by the healthcare provider (such as the physician) and the patient. While there is overlap in those factors and how they affect total healthcare costs, attention can be focused on those separate drivers of healthcare visit frequency and healthcare cost severity.
  - In personal lines, prior claims history is an important underwriting factor. It is common to use an indicator of whether or not the insured had a claim within a certain time period prior to the contract. Also, the number of claims incurred by the insured in previous periods has predictive power.
  - In homeowners insurance, in modeling potential loss frequency, the insurer could consider loss prevention measures that the homeowner has adopted, such as visible security systems. Separately, when modeling loss severity, the insurer would examine those factors that affect repair and replacement costs.
- **Databases.** Insurers may hold separate data files that suggest developing separate frequency and severity models. For example, a policyholder file is established when a policy is written. This file records much underwriting information about the insured(s), such as age, gender, and prior claims experience, policy information such as coverage, deductibles and limitations,

as well as any insurance claims event. A separate file, known as the “claims” file, records details of the claim against the insurer, including the amount. (There may also be a “payments” file that records the timing of the payments although we shall not deal with that here.) This recording process could then extend to insurers modeling the frequency and severity as separate processes.

- **Regulatory and Administrative.** Insurance is a highly regulated and monitored industry, given its importance in providing financial security to individuals and companies facing risk. As part of their duties, regulators routinely require the reporting of claims numbers as well as amounts. This may be due to the fact that there can be alternative definitions of an “amount,” e.g., paid versus incurred, and there is less potential error when reporting claim numbers. This continual monitoring helps ensure financial stability of these insurance companies.
- 

## 3.2 Basic Frequency Distributions

---

In this section, you learn how to:

- Determine quantities that summarize a distribution such as the distribution and survival function, as well as moments such as the mean and variance
  - Define and compute the moment and probability generating functions
  - Describe and understand relationships among three important frequency distributions: the binomial, Poisson, and negative binomial distributions
- 

In this section, we introduce the distributions that are commonly used in actuarial practice to model count data. The claim count random variable is denoted by  $N$ ; by its very nature it assumes only non-negative integer values. Hence the distributions below are all discrete distributions supported on the set of non-negative integers  $\{0, 1, \dots\}$ .

### 3.2.1 Foundations

Since  $N$  is a discrete random variable taking values in  $\{0, 1, \dots\}$ , the most natural full description of its distribution is through the specification of the probabilities with which it assumes each of the non-negative integer values. This leads us to the concept of the probability mass function (pmf) of  $N$ ,

denoted as  $p_N(\cdot)$  and defined as follows:

$$p_N(k) = \Pr(N = k), \quad \text{for } k = 0, 1, \dots$$

We note that there are alternate complete descriptions, or characterizations, of the distribution of  $N$ ; for example, the distribution function of  $N$  defined by  $F_N(x) = \Pr(N \leq x)$  and determined as:

$$F_N(x) = \begin{cases} \sum_{k=0}^{\lfloor x \rfloor} \Pr(N = k), & x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

In the above,  $\lfloor \cdot \rfloor$  denotes the floor function;  $\lfloor x \rfloor$  denotes the greatest integer less than or equal to  $x$ . This expression also suggests the descriptor *cumulative distribution function*, a commonly used alternative way of expressing the distribution function. We also note that the survival function of  $N$ , denoted by  $S_N(\cdot)$ , is defined as the ones'-complement of  $F_N(\cdot)$ , i.e.  $S_N(\cdot) = 1 - F_N(\cdot)$ . Clearly, the latter is another characterization of the distribution of  $N$ .

Often one is interested in quantifying a certain aspect of the distribution and not in its complete description. This is particularly useful when comparing distributions. A *center of location* of the distribution is one such aspect, and there are many different measures that are commonly used to quantify it. Of these, the mean is the most popular; the mean of  $N$ , denoted by  $\mu_N$ ,<sup>1</sup> is defined as

Another basic aspect of a distribution is its dispersion, and of the various measures of dispersion studied in the literature, the standard deviation is the

---

<sup>1</sup>For convenience, we have indexed  $\mu_N$  with the random variable  $N$  instead of  $F_N$  or  $p_N$ , even though it is solely a function of the distribution of the random variable.

$$\mu_N = \sum_{k=0}^{\infty} k p_N(k).$$

We note that  $\mu_N$  is the expected value of the random variable  $N$ , i.e.  $\mu_N = E[N]$ . This leads to a general class of measures, the moments of the distribution; the  $r$ -th raw moment of  $N$ , for  $r > 0$ , is defined as  $E[N^r]$  and denoted by  $\mu'_N(r)$ . We remark that the prime ' $'$  here does *not* denote differentiation. Rather, it is commonly used notation to distinguish a raw from a central moment, as will be introduced in Section 4.1.1. For  $r > 0$ , we have

$$\mu'_N(r) = E[N^r] = \sum_{k=0}^{\infty} k^r p_N(k).$$

We note that  $\mu'_N(\cdot)$  is a well-defined non-decreasing function taking values in  $[0, \infty]$ , as  $\Pr(N \in \{0, 1, \dots\}) = 1$ ; also, note that  $\mu_N = \mu'_N(1)$ . In the following, when we refer to a moment it will be implicit that it is finite unless mentioned otherwise.

most popular. Towards defining it, we first define the variance of  $N$ , denoted by  $\text{Var}[N]$ , as  $\text{Var}[N] = \mathbb{E}[(N - \mu_N)^2]$  when  $\mu_N$  is finite. By basic properties of the expected value of a random variable, we see that  $\text{Var}[N] = \mathbb{E}[N^2] - [\mathbb{E}(N)]^2$ . The standard deviation of  $N$ , denoted by  $\sigma_N$ , is defined as the square root of  $\text{Var}[N]$ . Note that the latter is well-defined as  $\text{Var}[N]$ , by its definition as the average squared deviation from the mean, is non-negative;  $\text{Var}[N]$  is denoted by  $\sigma_N^2$ . Note that these two measures take values in  $[0, \infty]$ .

### 3.2.2 Moment and Probability Generating Functions

Now we introduce two generating functions that are found to be useful when working with count variables. For a discrete random variable, the moment generating function (mgf) of  $N$ , denoted as  $M_N(\cdot)$ , is defined as

$$M_N(t) = \mathbb{E}[e^{tN}] = \sum_{k=0}^{\infty} e^{tk} p_N(k), \quad t \in \mathbb{R}.$$

We note that while  $M_N(\cdot)$  is well defined as it is the expectation of a non-negative random variable ( $e^{tN}$ ), it can assume the value  $\infty$ . Note that for a count random variable,  $M_N(\cdot)$  is finite valued on  $(-\infty, 0]$  with  $M_N(0) = 1$ . The following theorem, whose proof can be found in Billingsley (2008) (pages 285-6), encapsulates the reason for its name.

#### Theorem 3.1.

Let  $N$  be a count random variable such that  $\mathbb{E}[e^{t^*N}]$  is finite for some  $t^* > 0$ . We have the following:

- a. All moments of  $N$  are finite, *i.e.*

$$\mathbb{E}[N^r] < \infty, \quad r > 0.$$

- b. The *mgf* can be used to *generate* its moments as follows:

$$\frac{d^m}{dt^m} M_N(t) \Big|_{t=0} = \mathbb{E}[N^m], \quad m \geq 1.$$

- c. The *mgf*  $M_N(\cdot)$  characterizes the distribution; in other words it uniquely specifies the distribution.

Another reason that the *mgf* is very useful as a tool is that for two independent random variables  $X$  and  $Y$ , with their mgfs existing in a neighborhood of 0,

the *mgf* of  $X + Y$  is the product of their respective mgfs, that is,  $M_{X+Y}(t) = M_X(t)M_Y(t)$ , for small  $t$ .

A related generating function to the *mgf* is the probability generating function (*pgf*), and is a useful tool for random variables taking values in the non-negative integers. For a random variable  $N$ , by  $P_N(\cdot)$  we denote its *pgf* and we define it as follows<sup>2</sup>:

$$P_N(s) = \mathbb{E}[s^N], \quad s \geq 0.$$

It is straightforward to see that if the *mgf*  $M_N(\cdot)$  exists on  $(-\infty, t^*)$  then

$$P_N(s) = M_N(\log(s)), \quad s < e^{t^*}.$$

Moreover, if the *pgf* exists on an interval  $[0, s^*)$  with  $s^* > 1$ , then the *mgf*  $M_N(\cdot)$  exists on  $(-\infty, \log(s^*))$ , and hence uniquely specifies the distribution of  $N$  by [Theorem 3.1](#). (As a reminder, throughout this text we use *log* as the natural logarithm, not the base ten (common) logarithm or other version.) The following result for *pgf* is an analog of [Theorem 3.1](#), and in particular justifies its name.

### Theorem 3.2.

Let  $N$  be a count random variable such that  $\mathbb{E}(s^*)^N$  is finite for some  $s^* > 1$ . We have the following:

- a. All moments of  $N$  are finite, *i.e.*

$$\mathbb{E} N^r < \infty, \quad r \geq 0.$$

- b. The *pmf* of  $N$  can be derived from the *pgf* as follows:

$$p_N(m) = \begin{cases} P_N(0), & m = 0; \\ \left(\frac{1}{m!}\right) \frac{d^m}{ds^m} P_N(s) \Big|_{s=0}, & m \geq 1. \end{cases}$$

- c. The factorial moments of  $N$  can be derived as follows:

$$\left. \frac{d^m}{ds^m} P_N(s) \right|_{s=1} = \mathbb{E} \prod_{i=0}^{m-1} (N - i), \quad m \geq 1.$$

- d. The *pgf*  $P_N(\cdot)$  characterizes the distribution; in other words it uniquely specifies the distribution.

<sup>2</sup> $0^0 = 1$

### 3.2.3 Important Frequency Distributions

In this sub-section we study three important frequency distributions used in statistics, namely the binomial, the Poisson, and the negative binomial distributions. In the following, a risk denotes a unit covered by insurance. A risk could be an individual, a building, a company, or some other identifier for which insurance coverage is provided. For context, imagine an insurance data set containing the number of claims by risk or stratified in some other manner. The above mentioned distributions also happen to be the most commonly used in insurance practice for reasons, some of which we mention below.

- These distributions can be motivated by natural random experiments which are good approximations to real life processes from which many insurance data arise. Hence, not surprisingly, they together offer a reasonable fit to many insurance data sets of interest. The appropriateness of a particular distribution for the set of data can be determined using standard statistical methodologies, as we discuss later in this chapter.
- They provide a rich enough basis for generating other distributions that even better approximate or well cater to more real situations of interest to us.
  - The three distributions are either one-parameter or two-parameter distributions. In fitting to data, a parameter is assigned a particular value. The set of these distributions can be enlarged to their convex hulls by treating the parameter(s) as a random variable (or vector) with its own probability distribution, with this larger set of distributions offering greater flexibility. A simple example that is better addressed by such an enlargement is a portfolio of claims generated by insureds belonging to many different risk classes.
  - In insurance data, we may observe either a marginal or inordinate number of zeros, that is, zero claims by risk. When fitting to the data, a frequency distribution in its standard specification often fails to reasonably account for this occurrence. The natural modification of the above three distributions, however, accommodate this phenomenon well towards offering a better fit.
  - In insurance we are interested in total claims paid, whose distribution results from compounding the fitted frequency distribution with a severity distribution. These three distributions have properties that make it easy to work with the resulting aggregate severity distribution.

#### Binomial Distribution

We begin with the binomial distribution which arises from any finite sequence of identical and independent experiments with binary outcomes. The most canonical of such experiments is the (biased or unbiased) coin tossing experiment with the outcome being heads or tails. So if  $N$  denotes the number of

heads in a sequence of  $m$  independent coin tossing experiments with an identical coin which turns heads up with probability  $q$ , then the distribution of  $N$  is called the binomial distribution with parameters  $(m, q)$ , with  $m$  a positive integer and  $q \in [0, 1]$ . Note that when  $q = 0$  (resp.,  $q = 1$ ) then the distribution is degenerate with  $N = 0$  (resp.,  $N = m$ ) with probability 1. Clearly, its support when  $q \in (0, 1)$  equals  $\{0, 1, \dots, m\}$  with *pmf* given by <sup>3</sup>

$$p_k = \binom{m}{k} q^k (1 - q)^{m-k}, \quad k = 0, \dots, m.$$

where

$$\binom{m}{k} = \frac{m!}{k!(m-k)!}$$

The reason for its name is that the *pmf* takes values among the terms that arise from the binomial expansion of  $(q + (1-q))^m$ . This realization then leads to the the following expression for the *pgf* of the binomial distribution:

$$\begin{aligned} P_N(z) &= \sum_{k=0}^m z^k \binom{m}{k} q^k (1 - q)^{m-k} \\ &= \sum_{k=0}^m \binom{m}{k} (zq)^k (1 - q)^{m-k} \\ &= (qz + (1 - q))^m = (1 + q(z - 1))^m. \end{aligned}$$

Note that the above expression for the *pgf* confirms the fact that the binomial distribution is the  $m$ -convolution of the Bernoulli distribution, which is the binomial distribution with  $m = 1$  and *pgf*  $(1 + q(z - 1))$ . By “ $m$ -convolution,” we mean that we can write  $N$  as the sum of  $N_1, \dots, N_m$ . Here,  $N_i$  are iid Bernoulli variates. Also, note that the *mgf* of the binomial distribution is given by  $(1 + q(e^t - 1))^m$ .

The mean and variance of the binomial distribution can be found in a few different ways. To emphasize the key property that it is a  $m$ -convolution of the Bernoulli distribution, we derive below the moments using this property. We begin by observing that the Bernoulli distribution with parameter  $q$  assigns probability of  $q$  and  $1 - q$  to 1 and 0, respectively. So its mean equals  $q$  ( $= 0 \times (1 - q) + 1 \times q$ ); note that its raw second moment equals its mean as  $N^2 = N$  with probability 1. Using these two facts we see that the variance equals  $q(1 - q)$ . Moving on to the binomial distribution with parameters  $m$  and  $q$ , using the fact that it is the  $m$ -convolution of the Bernoulli distribution, we write  $N$  as the sum of  $N_1, \dots, N_m$ , where  $N_i$  are *iid* Bernoulli variates, as above. Now using the moments of Bernoulli and linearity of the expectation, we see that

$$E[N] = E \left[ \sum_{i=1}^m N_i \right] = \sum_{i=1}^m E[N_i] = mq.$$

---

<sup>3</sup>In the following we suppress the reference to  $N$  and denote the *pmf* by the sequence  $\{p_k\}_{k \geq 0}$ , instead of the function  $p_N(\cdot)$ .

Also, using the fact that the variance of the sum of independent random variables is the sum of their variances, we see that

$$\text{Var}[N] = \text{Var} \left[ \sum_{i=1}^m N_i \right] = \sum_{i=1}^m \text{Var}[N_i] = mq(1-q).$$

Alternate derivations of the above moments are suggested in the exercises. One important observation, especially from the point of view of applications, is that the mean is greater than the variance unless  $q = 0$ .

### Poisson Distribution

After the binomial distribution, the Poisson distribution (named after the French polymath Simeon Denis Poisson) is probably the most well known of discrete distributions. This is partly due to the fact that it arises naturally as the distribution of the count of the random occurrences of a type of event in a certain time period, if the rate of occurrences of such events is a constant. It also arises as the asymptotic limit of the binomial distribution with  $m \rightarrow \infty$  and  $mq \rightarrow \lambda$ .

The Poisson distribution is parametrized by a single parameter usually denoted by  $\lambda$  which takes values in  $(0, \infty)$ . Its *pmf* is given by

$$p_k = \frac{e^{-\lambda} \lambda^k}{k!}, k = 0, 1, \dots$$

It is easy to check that the above specifies a *pmf* as the terms are clearly non-negative, and that they sum to one follows from the infinite Taylor series expansion of  $e^\lambda$ . More generally, we can derive its *pgf*,  $P_N(\cdot)$ , as follows:

$$P_N(z) = \sum_{k=0}^{\infty} p_k z^k = \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^k z^k}{k!} = e^{-\lambda} e^{\lambda z} = e^{\lambda(z-1)}, \forall z \in \mathbb{R}.$$

From the above, we derive its *mgf* as follows:

$$M_N(t) = P_N(e^t) = e^{\lambda(e^t - 1)}, t \in \mathbb{R}.$$

Towards deriving its mean, we note that for the Poisson distribution

$$kp_k = \begin{cases} 0, & k = 0 \\ \lambda p_{k-1}, & k \geq 1. \end{cases}$$

This can be checked easily. In particular, this implies that

$$\mathbb{E}[N] = \sum_{k \geq 0} k p_k = \lambda \sum_{k \geq 1} p_{k-1} = \lambda \sum_{j \geq 0} p_j = \lambda.$$

In fact, more generally, using either a generalization of the above or using [Theorem 3.1](#), we see that

$$\mathbb{E} \prod_{i=0}^{m-1} (N - i) = \frac{d^m}{ds^m} P_N(s) \Big|_{s=1} = \lambda^m, \quad m \geq 1.$$

This, in particular, implies that

$$\text{Var}[N] = \mathbb{E}[N^2] - [\mathbb{E}(N)]^2 = \mathbb{E}[N(N-1)] + \mathbb{E}[N] - (\mathbb{E}[N])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Note that interestingly for the Poisson distribution  $\text{Var}[N] = \mathbb{E}[N]$ .

### Negative Binomial Distribution

The third important count distribution is the negative binomial distribution. Recall that the binomial distribution arose as the distribution of the number of *successes* in  $m$  independent repetitions of an experiment with binary outcomes. If we instead consider the number of *successes* until we observe the  $r$ -th *failure* in independent repetitions of an experiment with binary outcomes, then its distribution is a negative binomial distribution. A particular case, when  $r = 1$ , is the geometric distribution. However when  $r$  is not an integer, the above random experiment would not be applicable. In the following, we allow the parameter  $r$  to be any positive real number to then motivate the distribution more generally. To explain its name, we recall the binomial series, *i.e.*

$$(1+x)^s = 1 + sx + \frac{s(s-1)}{2!}x^2 + \dots, \quad s \in \mathbb{R}; |x| < 1.$$

If we define  $\binom{s}{k}$ , the generalized binomial coefficient, by

$$\binom{s}{k} = \frac{s(s-1)\cdots(s-k+1)}{k!},$$

then we have

$$(1+x)^s = \sum_{k=0}^{\infty} \binom{s}{k} x^k, \quad s \in \mathbb{R}; |x| < 1.$$

If we let  $s = -r$ , then we see that the above yields

$$(1-x)^{-r} = 1 + rx + \frac{(r+1)r}{2!}x^2 + \dots = \sum_{k=0}^{\infty} \binom{r+k-1}{k} x^k, \quad r \in \mathbb{R}; |x| < 1.$$

This implies that if we define  $p_k$  as

$$p_k = \binom{r+k-1}{k} \left( \frac{1}{1+\beta} \right)^r \left( \frac{\beta}{1+\beta} \right)^k, \quad k = 0, 1, \dots$$

for  $r > 0$  and  $\beta \geq 0$ , then it defines a valid *pmf*. Such defined distribution is called the negative binomial distribution with parameters  $(r, \beta)$  with  $r > 0$  and  $\beta \geq 0$ . Moreover, the binomial series also implies that the *pgf* of this distribution is given by

$$P_N(z) = (1 - \beta(z - 1))^{-r}, \quad |z| < 1 + \frac{1}{\beta}, \beta \geq 0.$$

The above implies that the *mgf* is given by

$$M_N(t) = (1 - \beta(e^t - 1))^{-r}, \quad t < \log\left(1 + \frac{1}{\beta}\right), \beta \geq 0.$$

We derive its moments using [Theorem 3.1](#) as follows:

$$\begin{aligned} E[N] &= M'(0) = r\beta e^t (1 - \beta(e^t - 1))^{-r-1} \Big|_{t=0} = r\beta; \\ E[N^2] &= M''(0) = [r\beta e^t (1 - \beta(e^t - 1))^{-r-1} + r(r+1)\beta^2 e^{2t} (1 - \beta(e^t - 1))^{-r-2}] \Big|_{t=0} \\ &= r\beta(1 + \beta) + r^2\beta^2; \\ \text{and } \text{Var}[N] &= E[N^2] - (E[N])^2 = r\beta(1 + \beta) + r^2\beta^2 - r^2\beta^2 = r\beta(1 + \beta) \end{aligned}$$

We note that when  $\beta > 0$ , we have  $\text{Var}[N] > E[N]$ . In other words, this distribution is overdispersed (relative to the Poisson); similarly, when  $q > 0$  the binomial distribution is said to be underdispersed (relative to the Poisson).

Finally, we observe that the Poisson distribution also emerges as a limit of negative binomial distributions. Towards establishing this, let  $\beta_r$  be such that as  $r$  approaches infinity  $r\beta_r$  approaches  $\lambda > 0$ . Then we see that the mgfs of negative binomial distributions with parameters  $(r, \beta_r)$  satisfies

$$\lim_{r \rightarrow 0} (1 - \beta_r(e^t - 1))^{-r} = \exp\{\lambda(e^t - 1)\},$$

with the right hand side of the above equation being the *mgf* of the Poisson distribution with parameter  $\lambda$ .<sup>4</sup>

### 3.3 The $(a, b, 0)$ Class

In this section, you learn how to:

- Define the  $(a,b,0)$  class of frequency distributions

---

<sup>4</sup>For the theoretical basis underlying the above argument, see [Billingsley \(2008\)](#).

- Discuss the importance of the recursive relationship underpinning this class of distributions
  - Identify conditions under which this general class reduces to each of the binomial, Poisson, and negative binomial distributions
- 

In the previous section we studied three distributions, namely the binomial, the Poisson and the negative binomial distributions. In the case of the Poisson, to derive its mean we used the fact that

$$kp_k = \lambda p_{k-1}, \quad k \geq 1,$$

which can be expressed equivalently as

$$\frac{p_k}{p_{k-1}} = \frac{\lambda}{k}, \quad k \geq 1.$$

Interestingly, we can similarly show that for the binomial distribution

$$\frac{p_k}{p_{k-1}} = \frac{-q}{1-q} + \left( \frac{(m+1)q}{1-q} \right) \frac{1}{k}, \quad k = 1, \dots, m,$$

and that for the negative binomial distribution

$$\frac{p_k}{p_{k-1}} = \frac{\beta}{1+\beta} + \left( \frac{(r-1)\beta}{1+\beta} \right) \frac{1}{k}, \quad k \geq 1.$$

The above relationships are all of the form

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k \geq 1; \tag{3.1}$$

this raises the question if there are any other distributions which satisfy this seemingly general recurrence relation. Note that the ratio on the left, the ratio of two probabilities, is non-negative.

---

**Snippet of Theory.** To begin with, let  $a < 0$ . In this case as  $k \rightarrow \infty$ ,  $(a + b/k) \rightarrow a < 0$ . It follows that if  $a < 0$  then  $b$  should satisfy  $b = -ka$ , for some  $k \geq 1$ . Any such pair  $(a, b)$  can be written as

$$\left( \frac{-q}{1-q}, \frac{(m+1)q}{1-q} \right), \quad q \in (0, 1), m \geq 1;$$

note that the case  $a < 0$  with  $a + b = 0$  yields the degenerate at 0 distribution which is the binomial distribution with  $q = 0$  and arbitrary  $m \geq 1$ .

In the case of  $a = 0$ , again by non-negativity of the ratio  $p_k/p_{k-1}$ , we have  $b \geq 0$ . If  $b = 0$  the distribution is degenerate at 0, which is a binomial with  $q = 0$  or a Poisson distribution with  $\lambda = 0$  or a negative binomial distribution with  $\beta = 0$ . If  $b > 0$ , then clearly such a distribution is a Poisson distribution with mean (*i.e.*  $\lambda$ ) equal to  $b$ , as presented at the beginning of this section.

In the case of  $a > 0$ , again by non-negativity of the ratio  $p_k/p_{k-1}$ , we have  $a+b/k \geq 0$  for all  $k \geq 1$ . The most stringent of these is the inequality  $a+b \geq 0$ . Note that  $a + b = 0$  again results in degeneracy at 0; excluding this case we have  $a + b > 0$  or equivalently  $b = (r - 1)a$  with  $r > 0$ . Some algebra easily yields the following expression for  $p_k$ :

$$p_k = \binom{r+k-1}{k} p_0 a^k, \quad k = 1, 2, \dots$$

The above series converges for  $a < 1$  when  $r > 0$ , with the sum given by  $p_0 \cdot ((1-a)^{(-r)} - 1)$ . Hence, equating the latter to  $1-p_0$  we get  $p_0 = (1-a)^{(r)}$ . So in this case the pair  $(a, b)$  is of the form  $(a, (r-1)a)$ , for  $r > 0$  and  $0 < a < 1$ ; since an equivalent parametrization is  $(\beta/(1+\beta), (r-1)\beta/(1+\beta))$ , for  $r > 0$  and  $\beta > 0$ , we see from above that such distributions are negative binomial distributions.

From the above development we see that not only does the recurrence (3.1) tie these three distributions together, but also it characterizes them. For this reason these three distributions are collectively referred to in the actuarial literature as  $(a,b,0)$  class of distributions, with 0 referring to the starting point of the recurrence. Note that the value of  $p_0$  is implied by  $(a, b)$  since the probabilities have to sum to one. Of course, (3.1) as a recurrence relation for  $p_k$  makes the computation of the *pmf* efficient by removing redundancies. Later, we will see that it does so even in the case of compound distributions with the frequency distribution belonging to the  $(a, b, 0)$  class - this fact is the more important motivating reason to study these three distributions from this viewpoint.

**Example 3.3.1.** A discrete probability distribution has the following properties

$$\begin{aligned} p_k &= c \left(1 + \frac{2}{k}\right) p_{k-1} \quad k = 1, 2, 3, \dots \\ p_1 &= \frac{9}{256} \end{aligned}$$

Determine the expected value of this discrete random variable.

**Example Solution.** Since the *pmf* satisfies the  $(a, b, 0)$  recurrence relation we know that the underlying distribution is one among the binomial, Poisson, and negative binomial distributions. Since the ratio of the parameters (\*i.e.\*  $b/a$ ) equals 2, we know that it is negative binomial and that  $r = 3$ . Moreover, since for a negative binomial  $p_1 = r(1 + \beta)^{-(r+1)}\beta$ , we have

$$\begin{aligned}\frac{9}{256} &= 3 \frac{\beta}{(1 + \beta)^4} \\ \Rightarrow \quad \frac{3}{(1 + 3)^4} &= \frac{\beta}{(1 + \beta)^4} \\ \Rightarrow \quad \beta &= 3.\end{aligned}$$

Finally, since the mean of a negative binomial is  $r\beta$  we have the mean of the given distribution equals 9.

## 3.4 Estimating Frequency Distributions

---



---

In this section, you learn how to:

- Define a likelihood for a sample of observations from a discrete distribution
  - Define the maximum likelihood estimator for a random sample of observations from a discrete distribution
  - Calculate the maximum likelihood estimator for the binomial, Poisson, and negative binomial distributions
- 

### 3.4.1 Parameter Estimation

In Section 3.2 we introduced three distributions of importance in modeling various types of count data arising from insurance. Let us now suppose that we have a set of count data to which we wish to fit a distribution, and that we have determined that one of these  $(a, b, 0)$  distributions is more appropriate than the others. Since each one of these forms a class of distributions if we allow its parameter(s) to take any permissible value, there remains the task of determining the **best** value of the parameter(s) for the data at hand. This is a statistical point estimation problem, and in parametric inference problems the statistical inference paradigm of *maximum likelihood* usually yields efficient

estimators. In this section we describe this paradigm and derive the maximum likelihood estimators.

Let us suppose that we observe the independent and identically distributed, *iid*, random variables  $X_1, X_2, \dots, X_n$  from a distribution with pmf  $p_\theta$ , where  $\theta$  is a vector of parameters and an unknown value in the parameter space  $\Theta \subseteq \mathbb{R}^d$ . For example, in the case of the Poisson distribution, there is a single parameter so that  $d = 1$  and

$$p_\theta(x) = e^{-\theta} \frac{\theta^x}{x!}, \quad x = 0, 1, \dots,$$

with  $\theta > 0$ . In the case of the binomial distribution we have

$$p_\theta(x) = \binom{m}{x} q^x (1-q)^{m-x}, \quad x = 0, 1, \dots, m.$$

For some applications, we can view  $m$  as a parameter and so take  $d = 2$  so that  $\theta = (m, q) \in \{0, 1, 2, \dots\} \times [0, 1]$ .

Let us suppose that the observations are  $x_1, \dots, x_n$ , observed values of the random sample  $X_1, X_2, \dots, X_n$  presented earlier. In this case, the probability of observing this sample from  $p_\theta$  equals

$$\prod_{i=1}^n p_\theta(x_i).$$

The above, denoted by  $L(\theta)$ , viewed as a function of  $\theta$ , is called the *likelihood*. Note that we suppressed its dependence on the data, to emphasize that we are viewing it as a function of the parameter vector. For example, in the case of the Poisson distribution we have

$$L(\lambda) = e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i} \left( \prod_{i=1}^n x_i! \right)^{-1}.$$

In the case of the binomial distribution we have

$$L(m, q) = \left( \prod_{i=1}^n \binom{m}{x_i} \right) q^{\sum_{i=1}^n x_i} (1-q)^{nm - \sum_{i=1}^n x_i}.$$

The maximum likelihood estimator (mle) for  $\theta$  is any maximizer of the likelihood; in a sense the *mle* chooses the set of parameter values that best explains the observed observations. Appendix Section ?? reviews the foundations of maximum likelihood estimation with more mathematical details in Appendix Chapter ??.

**Special Case: Three Bernoulli Outcomes.** To illustrate, consider a sample

of size  $n = 3$  from a Bernoulli distribution (binomial with  $m = 1$ ) with values  $0, 1, 0$ . The likelihood in this case is easily checked to equal

$$L(q) = q(1 - q)^2,$$

and the plot of the likelihood is given in Figure 3.1. As shown in the plot, the maximum value of the likelihood equals  $4/27$  and is attained at  $q = 1/3$ , and hence the maximum likelihood estimate for  $q$  is  $1/3$  for the given sample. In this case one can resort to algebra to show that

$$q(1 - q)^2 = \left(q - \frac{1}{3}\right)^2 \left(q - \frac{4}{3}\right) + \frac{4}{27},$$

and conclude that the maximum equals  $4/27$ , and is attained at  $q = 1/3$  (using the fact that the first term is non-positive in the interval  $[0, 1]$ ).

But as is apparent, this way of deriving the *mle* using algebra does not generalize. In general, one resorts to calculus to derive the *mle* - note that for some likelihoods one may have to resort to other optimization methods, especially when the likelihood has many local extrema. It is customary to equivalently maximize the logarithm of the likelihood<sup>5</sup>  $L(\cdot)$ , denoted by  $l(\cdot)$ , and look at the set of zeros of its first derivative<sup>6</sup>  $l'(\cdot)$ . In the case of the above likelihood,  $l(q) = \log(q) + 2\log(1 - q)$ , and

$$l'(q) = \frac{d}{dq} l(q) = \frac{1}{q} - \frac{2}{1 - q}.$$

The unique zero of  $l'(\cdot)$  equals  $1/3$ , and since  $l''(\cdot)$  is negative, we have  $1/3$  is the unique maximizer of the likelihood and hence its maximum likelihood estimate.

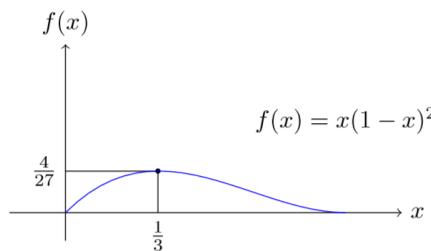


FIGURE 3.1: Likelihood of a  $(0, 1, 0)$  3-sample from Bernoulli

---

<sup>5</sup>The set of maximizers of  $L(\cdot)$  are the same as the set of maximizers of any strictly increasing function of  $L(\cdot)$ , and hence the same as those for  $l(\cdot)$ .

<sup>6</sup>A slight benefit of working with  $l(\cdot)$  is that constant terms in  $L(\cdot)$  do not appear in  $l'(\cdot)$  whereas they do in  $L'(\cdot)$ .

### 3.4.2 Frequency Distributions MLE

In the following, we derive the maximum likelihood estimator, *mle*, for the three members of the  $(a, b, 0)$  class. We begin by summarizing the discussion above. In the setting of observing *iid*, independent and identically distributed, random variables  $X_1, X_2, \dots, X_n$  from a distribution with pmf  $p_\theta$ , where  $\theta$  takes an unknown value in  $\Theta \subseteq \mathbb{R}^d$ , the likelihood  $L(\cdot)$ , a function on  $\Theta$  is defined as

$$L(\theta) = \prod_{i=1}^n p_\theta(x_i),$$

where  $x_1, \dots, x_n$  are the observed values. The *mle* of  $\theta$ , denoted as  $\hat{\theta}_{\text{MLE}}$ , is a function which maps the observations to an element of the set of maximizers of  $L(\cdot)$ , namely

$$\{\theta | L(\theta) = \max_{\eta \in \Theta} L(\eta)\}.$$

Note the above set is a function of the observations, even though this dependence is not made explicit. In the case of the three distributions that we study, and quite generally, the above set is a singleton with probability tending to one (with increasing sample size). In other words, for many commonly used distributions and when the sample size is large, the likelihood estimate is uniquely defined with high probability.

In the following, we assume that we have observed  $n$  *iid* random variables  $X_1, X_2, \dots, X_n$  from the distribution under consideration, even though the parametric value is unknown. Also,  $x_1, x_2, \dots, x_n$  will denote the observed values. We note that in the case of count data, and data from discrete distributions in general, the likelihood can alternately be represented as

$$L(\theta) = \prod_{k \geq 0} (p_\theta(k))^{m_k},$$

where  $m_k$  is the number of observations equal to  $k$ . Mathematically, we have

$$m_k = |\{i | x_i = k, 1 \leq i \leq n\}| = \sum_{i=1}^n I(x_i = k), \quad k \geq 0.$$

Note that this transformation retains all of the data, compiling it in a streamlined manner. For large  $n$  it leads to compression of the data in the sense of *sufficiency*. Below, we present expressions for the *mle* in terms of  $\{m_k\}_{k \geq 1}$  as well.

**Special Case: Poisson Distribution.** In this case, as noted above, the likelihood is given by

$$L(\lambda) = \left( \prod_{i=1}^n x_i! \right)^{-1} e^{-n\lambda} \lambda^{\sum_{i=1}^n x_i}.$$

Taking logarithms, the log-likelihood is

$$l(\lambda) = - \sum_{i=1}^n \log(x_i!) - n\lambda + \log(\lambda) \cdot \sum_{i=1}^n x_i.$$

Taking a derivative, we have

$$l'(\lambda) = -n + \frac{1}{\lambda} \sum_{i=1}^n x_i.$$

In evaluating  $l''(\lambda)$ , when  $\sum_{i=1}^n x_i > 0$ ,  $l'' < 0$ . Consequently, the maximum is attained at the sample mean,  $\bar{x}$ , presented below. When  $\sum_{i=1}^n x_i = 0$ , the likelihood is a decreasing function and hence the maximum is attained at the least possible parameter value; this results in the maximum likelihood estimate being zero. Hence, we have

$$\bar{x} = \hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Note that the sample mean can be computed also as

$$\bar{x} = \frac{1}{n} \sum_{k \geq 1} k \cdot m_k .$$

It is noteworthy that in the case of the Poisson, the exact distribution of  $\hat{\lambda}_{MLE}$  is available in closed form - it is a scaled Poisson - when the underlying distribution is a Poisson. This is so as the sum of independent Poisson random variables is a Poisson as well. Of course, for large sample size one can use the ordinary Central Limit Theorem (CLT) to derive a normal approximation. Note that the latter approximation holds even if the underlying distribution is any distribution with a finite second moment.

**Special Case: Binomial Distribution with known  $m$ .** Unlike the case of the Poisson distribution, the parameter space in the case of the binomial is 2-dimensional. Hence the optimization problem is a bit more challenging. We first discuss the case where  $m$  is taken to be known - this is not a realistic assumption in insurance applications but is appropriate in circumstances where we are observing  $m$  iid binary outcomes with unknown probabilities.

We begin by observing that the likelihood is given by

$$L(m, q) = \left( \prod_{i=1}^n \binom{m}{x_i} \right) q^{\sum_{i=1}^n x_i} (1-q)^{nm - \sum_{i=1}^n x_i}.$$

Taking logarithms, the log-likelihood is

$$\begin{aligned} l(m, q) &= \sum_{i=1}^n \log \left( \binom{m}{x_i} \right) + (\sum_{i=1}^n x_i) \log(q) \\ &\quad + (nm - \sum_{i=1}^n x_i) \log(1-q) \\ &= \sum_{i=1}^n \log \left( \binom{m}{x_i} \right) + n\bar{x} \log(q) + n(m-\bar{x}) \log(1-q), \end{aligned}$$

where  $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ . Since we have assumed that  $m$  is known, maximizing  $l(m, q)$  with respect to  $q$  involves taking the first differential and equating to zero:

$$\frac{dl(m, q)}{dq} = \frac{n\bar{x}}{q} - \frac{n(m-\bar{x})}{1-q} = 0,$$

which implies that

$$\hat{q}_{MLE} = \frac{\bar{x}}{m}.$$

**Special Case: Binomial Distribution with unknown  $m$ .** Note that since  $m$  takes only non-negative integer values, we cannot use multivariate calculus to find the optimal values. Nevertheless, we can use single variable calculus to show that

$$\hat{q}_{MLE} \times \hat{m}_{MLE} = \bar{x}. \quad (3.2)$$


---

Towards this we note that for a fixed value of  $m$ ,

$$\frac{\delta}{\delta q} l(m, q) = \frac{n\bar{x}}{q} - \frac{n(m-\bar{x})}{1-q},$$

and that

$$\frac{\delta^2}{\delta q^2} l(m, q) = -\frac{n\bar{x}}{q^2} + \frac{n(m-\bar{x})}{(1-q)^2} \leq 0.$$

The above implies that for any fixed value of  $m$ , the maximizing value of  $q$  satisfies

$$mq = \bar{x},$$

and hence we establish equation (3.2).

---

With equation (3.2), the above reduces the task to the search for  $\hat{m}_{MLE}$ , which is a maximizer of

$$L \left( m, \frac{\bar{x}}{m} \right). \quad (3.3)$$

Note the likelihood would be zero for values of  $m$  smaller than  $\max_{1 \leq i \leq n} x_i$ , and hence  $\hat{m}_{MLE} \geq \max_{1 \leq i \leq n} x_i$ .

---

Towards specifying an algorithm to compute  $\hat{m}_{MLE}$ , we first point out that for some data sets  $\hat{m}_{MLE}$  could equal  $\infty$ , indicating that a Poisson distribution would render a better fit than any binomial distribution. This is so as the binomial distribution with parameters  $(m, \bar{x}/m)$  approaches the Poisson distribution with parameter  $\bar{x}$  with  $m$  approaching infinity. The fact that some data sets **prefer** a Poisson distribution should not be surprising since in the above sense the set of Poisson distribution is on the boundary of the set of binomial distributions. Interestingly, in [Olkin et al. \(1981\)](#) they show that if the sample mean is less than or equal to the sample variance then  $\hat{m}_{MLE} = \infty$ ; otherwise, there exists a finite  $m$  that maximizes equation (3.3).

---

In Figure 3.2 below we display the plot of  $L(m, \bar{x}/m)$  for three different samples of size 5; they differ only in the value of the sample maximum. The first sample of  $(2, 2, 2, 4, 5)$  has the ratio of sample mean to sample variance greater than 1 (1.875), the second sample of  $(2, 2, 2, 4, 6)$  has the ratio equal to 1.25 which is closer to 1, and the third sample of  $(2, 2, 2, 4, 7)$  has the ratio less than 1 (0.885). For these three samples, as shown in Figure 3.2,  $\hat{m}_{MLE}$  equals 7, 18 and  $\infty$ , respectively. Note that the limiting value of  $L(m, \bar{x}/m)$  as  $m$  approaches infinity equals

$$\left( \prod_{i=1}^n x_i! \right)^{-1} \exp(-n\bar{x}) (\bar{x})^{n\bar{x}}. \quad (3.4)$$

Also, note that Figure 3.2 shows that the *mle* of  $m$  is non-robust, *i.e.* changes in a small proportion of the data set can cause large changes in the estimator.

The above discussion suggests the following simple algorithm:

- *Step 1.* If the sample mean is less than or equal to the sample variance, then set  $\hat{m}_{MLE} = \infty$ . The *mle* suggested distribution is a Poisson distribution with  $\hat{\lambda} = \bar{x}$ .
- *Step 2.* If the sample mean is greater than the sample variance, then compute  $L(m, \bar{x}/m)$  for  $m$  values greater than or equal to the sample maximum until  $L(m, \bar{x}/m)$  is close to the value of the Poisson likelihood given in (3.4). The value of  $m$  that corresponds to the maximum value of  $L(m, \bar{x}/m)$  among those computed equals  $\hat{m}_{MLE}$ .

We note that if the underlying distribution is the binomial distribution with parameters  $(m, q)$  (with  $q > 0$ ) then  $\hat{m}_{MLE}$  equals  $m$  for large sample sizes. Also,  $\hat{q}_{MLE}$  will have an asymptotically normal distribution and converge with probability one to  $q$ .

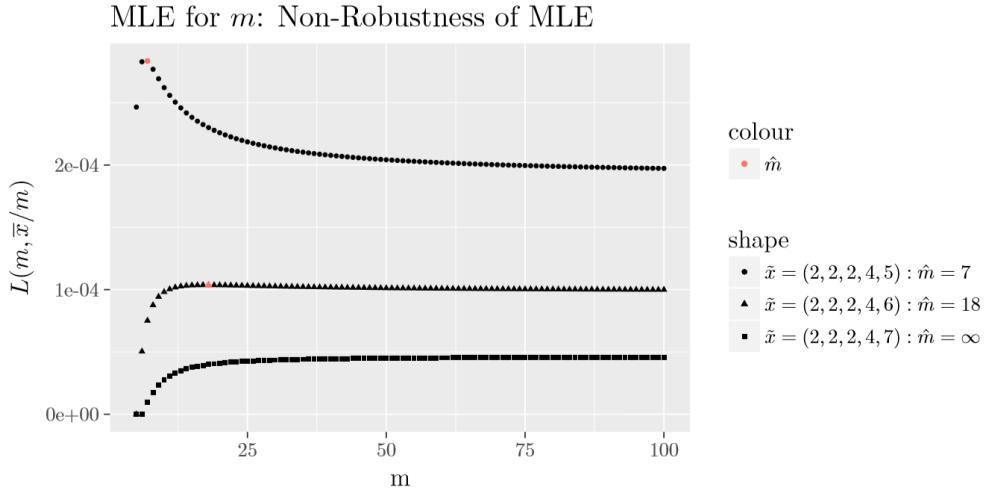


FIGURE 3.2: Plot of  $L(m, \bar{x}/m)$  for a Binomial Distribution

**Special Case: Negative Binomial Distribution.** The case of the negative binomial distribution is similar to that of the binomial distribution in the sense that we have two parameters and the *mles* are not available in closed form. A difference between them is that unlike the binomial parameter  $m$  which takes positive integer values, the parameter  $r$  of the negative binomial can assume any positive real value. This makes the optimization problem a tad more complex. We begin by observing that the likelihood can be expressed in the following form:

$$L(r, \beta) = \left( \prod_{i=1}^n \binom{r + x_i - 1}{x_i} \binom{r + x_i - 1}{x_i} \right) (1 + \beta)^{-n(r + \bar{x})} \beta^{n\bar{x}}.$$

The above implies that log-likelihood is given by

$$l(r, \beta) = \sum_{i=1}^n \log \binom{r + x_i - 1}{x_i} - n(r + \bar{x}) \log(1 + \beta) + n\bar{x} \log \beta,$$

and hence

$$\frac{\delta}{\delta \beta} l(r, \beta) = -\frac{n(r + \bar{x})}{1 + \beta} + \frac{n\bar{x}}{\beta}.$$

Equating the above to zero, we get

$$\hat{r}_{MLE} \times \hat{\beta}_{MLE} = \bar{x}.$$

The above reduces the two dimensional optimization problem to a one-dimensional problem - we need to maximize

$$l(r, \bar{x}/r) = \sum_{i=1}^n \log \binom{r + x_i - 1}{x_i} - n(r + \bar{x}) \log(1 + \bar{x}/r) + n\bar{x} \log(\bar{x}/r),$$

with respect to  $r$ , with the maximizing  $r$  being its *mle* and  $\hat{\beta}_{MLE} = \bar{x}/\hat{r}_{MLE}$ . In Levin et al. (1977) it is shown that if the sample variance is greater than the sample mean then there exists a unique  $r > 0$  that maximizes  $l(r, \bar{x}/r)$  and hence a unique *mle* for  $r$  and  $\beta$ . Also, they show that if  $\hat{\sigma}^2 \leq \bar{x}$ , then the negative binomial likelihood will be dominated by the Poisson likelihood with  $\hat{\lambda} = \bar{x}$ . In other words, a Poisson distribution offers a better fit to the data. The guarantee in the case of  $\hat{\sigma}^2 > \hat{\mu}$  permits us to use some algorithm to maximize  $l(r, \bar{x}/r)$ . Towards an alternate method of computing the likelihood, we note that

$$\begin{aligned} l(r, \bar{x}/r) &= \sum_{i=1}^n \sum_{j=1}^{x_i} \log(r - 1 + j) - \sum_{i=1}^n \log(x_i!) \\ &\quad - n(r + \bar{x}) \log(r + \bar{x}) + nr \log(r) + n\bar{x} \log(\bar{x}), \end{aligned}$$

which yields

$$\left(\frac{1}{n}\right) \frac{\delta}{\delta r} l(r, \bar{x}/r) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{x_i} \frac{1}{r - 1 + j} - \log(r + \bar{x}) + \log(r).$$

We note that, in the above expressions for the terms involving a double summation, the inner sum equals zero if  $x_i = 0$ . The *maximum likelihood estimate* for  $r$  is a root of the last expression and we can use a root finding algorithm to compute it. Also, we have

$$\left(\frac{1}{n}\right) \frac{\delta^2}{\delta r^2} l(r, \bar{x}/r) = \frac{\bar{x}}{r(r + \bar{x})} - \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{x_i} \frac{1}{(r - 1 + j)^2}.$$

A simple but quickly converging iterative root finding algorithm is the Newton's method, which incidentally the Babylonians are believed to have used for computing square roots. Under this method, an initial approximation is selected for the root and new approximations for the root are successively generated until convergence. Applying the Newton's method to our problem results in the following algorithm:

*Step i.* Choose an approximate solution, say  $r_0$ . Set  $k$  to 0.

*Step ii.* Define  $r_{k+1}$  as

$$r_{k+1} = r_k - \frac{\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{x_i} \frac{1}{r_k - 1 + j} - \log(r_k + \bar{x}) + \log(r_k)}{\frac{\bar{x}}{r_k(r_k + \bar{x})} - \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{x_i} \frac{1}{(r_k - 1 + j)^2}}$$

*Step iii.* If  $r_{k+1} \sim r_k$ , then report  $r_{k+1}$  as maximum likelihood estimate; else increment  $k$  by 1 and repeat *Step ii*.

For example, we simulated a 5 observation sample of 41, 49, 40, 27, 23 from the negative binomial with parameters  $r = 10$  and  $\beta = 5$ . Choosing the starting value of  $r$  such that

$$r\beta = \hat{\mu} \quad \text{and} \quad r\beta(1 + \beta) = \hat{\sigma}^2$$

where  $\hat{\mu}$  represents the estimated mean and  $\hat{\sigma}^2$  is the estimated variance. This leads to the starting value for  $r$  of 23.14286. The iterates of  $r$  from the Newton's method are

$$21.39627, 21.60287, 21.60647, 21.60647;$$

the rapid convergence seen above is typical of the Newton's method. Hence in this example,  $\hat{r}_{MLE} \sim 21.60647$  and  $\hat{\beta}_{MLE} = 1.66616$ .

---

```
Newton <- function(x, abserr) {
  mu <- mean(x)
  sigma2 <- mean(x^2) - mu^2
  r <- mu^2/(sigma2 - mu)
  b <- TRUE
  iter <- 0
  while (b) {
    tr <- r
    m1 <- mean(c(x[x == 0], sapply(x[x > 0], function(z) {
      sum(1/(tr:(tr - 1 + z)))
    })))
    m2 <- mean(c(x[x == 0], sapply(x[x > 0], function(z) {
      sum(1/(tr:(tr - 1 + z))^2)
    })))
    r <- tr - (m1 - log(1 + mu/tr))/(mu/(tr * (tr + mu)) - m2)
    b <- !(abs(tr - r) < abserr)
    iter <- iter + 1
  }
  c(r, iter)
}
```

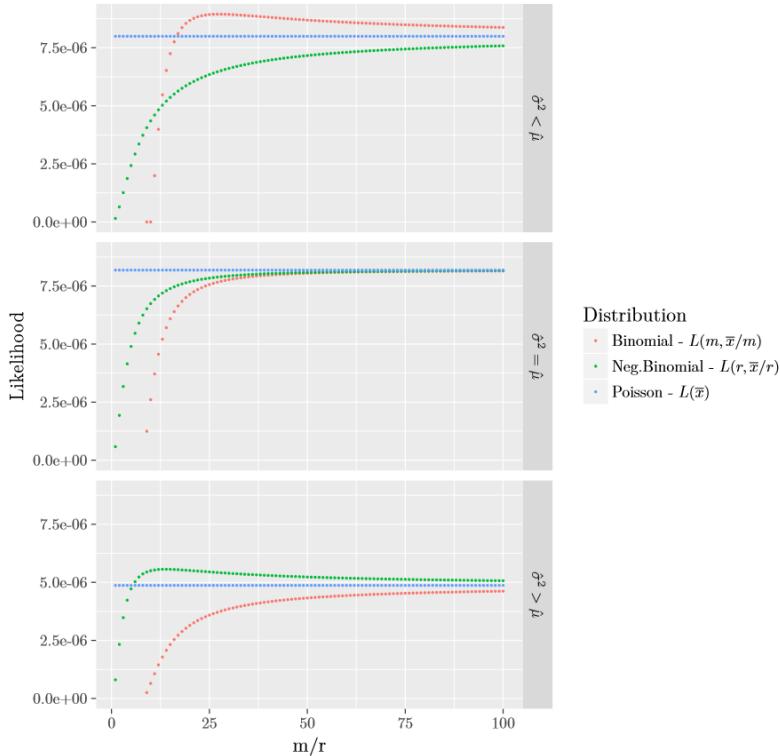
---

To summarize our discussion of MLE for the  $(a, b, 0)$  class of distributions, in Figure 3.3 below we plot the maximum value of the Poisson likelihood,  $L(m, \bar{x}/m)$  for the binomial, and  $L(r, \bar{x}/r)$  for the negative binomial, for the three samples of size 5 given in Table 3.1. The data was constructed to cover the three orderings of the sample mean and variance. As shown in the Figure 3.3, and supported by theory, if  $\hat{\mu} < \hat{\sigma}^2$  then the negative binomial results

in a higher maximum likelihood value; if  $\hat{\mu} = \hat{\sigma}^2$  the Poisson has the highest likelihood value; and finally in the case that  $\hat{\mu} > \hat{\sigma}^2$  the binomial gives a better fit than the others. So before fitting a frequency data with an  $(a, b, 0)$  distribution, it is best to start with examining the ordering of  $\hat{\mu}$  and  $\hat{\sigma}^2$ . We again emphasize that the Poisson is on the **boundary** of the negative binomial and binomial distributions. So in the case that  $\hat{\mu} \geq \hat{\sigma}^2$  ( $\hat{\mu} \leq \hat{\sigma}^2$ , resp.) the Poisson yields a better fit than the negative binomial (binomial, resp.), which is indicated by  $\hat{r} = \infty$  ( $\hat{m} = \infty$ , respectively).

Table 3.1. Three Samples of Size 5

Data	Mean ( $\hat{\mu}$ )	Variance ( $\hat{\sigma}^2$ )
(2, 3, 6, 8, 9)	5.60	7.44
(2, 5, 6, 8, 9)	6	6
(4, 7, 8, 10, 11)	8	6

FIGURE 3.3: Plot of  $(a, b, 0)$  Partially Maximized Likelihoods

---

### 3.5 Other Frequency Distributions

---

In this section, you learn how to:

- Define the  $(a,b,1)$  class of frequency distributions and discuss the importance of the recursive relationship underpinning this class of distributions
- Interpret zero truncated and modified versions of the binomial, Poisson, and negative binomial distributions

- Compute probabilities using the recursive relationship
- 

In the previous sections, we discussed three distributions with supports contained in the set of non-negative integers, which well cater to many insurance applications. Moreover, typically by allowing the parameters to be a function of known (to the insurer) explanatory variables such as age, sex, geographic location (territory), and so forth, these distributions allow us to explain claim probabilities in terms of these variables. The field of statistical study that studies such models is known as regression analysis - it is an important topic of actuarial interest that we will not pursue in this book; see [Frees \(2009\)](#).

There are clearly infinitely many other count distributions, and more importantly the above distributions by themselves do not cater to all practical needs. In particular, one feature of some insurance data is that the proportion of zero counts can be out of place with the proportion of other counts to be explainable by the above distributions. In the following we modify the above distributions to allow for arbitrary probability for zero count irrespective of the assignment of relative probabilities for the other counts. Another feature of a data set which is naturally comprised of homogeneous subsets is that while the above distributions may provide good fits to each subset, they may fail to do so to the whole data set. Later we naturally extend the  $(a, b, 0)$  distributions to be able to cater to, in particular, such data sets.

### 3.5.1 Zero Truncation or Modification

Let us suppose that we are looking at auto insurance policies which appear in a database of auto claims made in a certain period. If one is to study the number of claims that these policies have made during this period, then clearly the distribution has to assign a probability of zero to the count variable assuming the value zero. In other words, by restricting attention to count data from policies in the database of claims, we have in a sense zero-truncated the count data of all policies. In personal lines (like auto), policyholders may not want to report that first claim because of fear that it may increase future insurance rates - this behavior inflates the proportion of zero counts. Examples such as the latter modify the proportion of zero counts. Interestingly, natural modifications of the three distributions considered above are able to provide good fits to zero-modified/truncated data sets arising in insurance.

As presented below, we modify the probability assigned to zero count by the  $(a, b, 0)$  class while maintaining the relative probabilities assigned to non-zero counts - zero modification. Note that since the  $(a, b, 0)$  class of distributions satisfies the recurrence (3.1), maintaining relative probabilities of non-zero

counts implies that recurrence (3.1) is satisfied for  $k \geq 2$ . This leads to the definition of the following class of distributions.

**Definition.** A count distribution is a member of the  $(a, b, 1)$  class if for some constants  $a$  and  $b$  the probabilities  $p_k$  satisfy

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k \geq 2. \quad (3.5)$$

Note that since the recursion starts with  $p_1$ , and not  $p_0$ , we refer to this super-class of  $(a, b, 0)$  distributions by  $(a, b, 1)$ . To understand this class, recall that each valid pair of values for  $a$  and  $b$  of the  $(a, b, 0)$  class corresponds to a unique vector of probabilities  $\{p_k\}_{k \geq 0}$ . If we now look at the probability vector  $\{\tilde{p}_k\}_{k \geq 0}$  given by

$$\tilde{p}_k = \frac{1 - \tilde{p}_0}{1 - p_0} \cdot p_k, \quad k \geq 1,$$

where  $\tilde{p}_0 \in [0, 1]$  is arbitrarily chosen, then since the relative probabilities for positive values according to  $\{p_k\}_{k \geq 0}$  and  $\{\tilde{p}_k\}_{k \geq 0}$  are the same, we have  $\{\tilde{p}_k\}_{k \geq 0}$  satisfies recurrence (3.5). This, in particular, shows that the class of  $(a, b, 1)$  distributions is strictly wider than that of  $(a, b, 0)$ .

In the above, we started with a pair of values for  $a$  and  $b$  that led to a valid  $(a, b, 0)$  distribution, and then looked at the  $(a, b, 1)$  distributions that corresponded to this  $(a, b, 0)$  distribution. We now argue that the  $(a, b, 1)$  class allows for a larger set of permissible distributions for  $a$  and  $b$  than the  $(a, b, 0)$  class. Recall from Section 3.3 that in the case of  $a < 0$  we did not use the fact that the recurrence (3.1) started at  $k = 1$ , and hence the set of pairs  $(a, b)$  with  $a < 0$  that are permissible for the  $(a, b, 0)$  class is identical to those that are permissible for the  $(a, b, 1)$  class. The same conclusion is easily drawn for pairs with  $a = 0$ . In the case that  $a > 0$ , instead of the constraint  $a + b > 0$  for the  $(a, b, 0)$  class we now have the weaker constraint of  $a + b/2 > 0$  for the  $(a, b, 1)$  class. With the parametrization  $b = (r - 1)a$  as used in Section 3.3, instead of  $r > 0$  we now have the weaker constraint of  $r > -1$ . In particular, we see that while zero modifying a  $(a, b, 0)$  distribution leads to a distribution in the  $(a, b, 1)$  class, the conclusion does not hold in the other direction.

Zero modification of a count distribution  $F$  such that it assigns zero probability to zero count is called a zero truncation of  $F$ . Hence, the zero truncated version of probabilities  $\{p_k\}_{k \geq 0}$  is given by

$$\tilde{p}_k = \begin{cases} 0, & k = 0; \\ \frac{p_k}{1 - p_0}, & k \geq 1. \end{cases}$$

In particular, we have that a zero modification of a count distribution  $\{p_k^T\}_{k \geq 0}$ , denoted by  $\{p_k^M\}_{k \geq 0}$ , can be written as a convex combination of the degenerate distribution at 0 and the zero truncation of  $\{p_k\}_{k \geq 0}$ , denoted by  $\{p_k^T\}_{k \geq 0}$ . That is we have

$$p_k^M = p_0^M \cdot \delta_0(k) + (1 - p_0^M) \cdot p_k^T, \quad k \geq 0.$$

**Example 3.5.1. Zero Truncated/Modified Poisson.** Consider a Poisson distribution with parameter  $\lambda = 2$ . Calculate  $p_k, k = 0, 1, 2, 3$ , for the usual (unmodified), truncated and a modified version with ( $p_0^M = 0.6$ ).

**Example Solution.** For the Poisson distribution as a member of the  $(a, b, 0)$  class, we have  $a = 0$  and  $b = \lambda = 2$ . Thus, we may use the recursion  $p_k = \lambda p_{k-1}/k = 2p_{k-1}/k$  for each type, after determining starting probabilities. The calculation of probabilities for  $k \leq 3$  is shown in the following table.

Table. \*\*Calculation of Probabilities for\*\*  $k \leq 3$

$k$	$p_k$	$p_k^T$	$p_k^M$
0	$p_0 = e^{-\lambda} = 0.135335$	0	0.6
1	$p_1 = p_0(0 + \frac{\lambda}{1}) = 0.27067$	$\frac{p_1}{1-p_0} = 0.313035$	$\frac{1-p_0^M}{1-p_0} p_1 = 0.125214$
2	$p_2 = p_1(\frac{\lambda}{2}) = 0.27067$	$p_2^T = p_1^T(\frac{\lambda}{2}) = 0.313035$	$p_2^M = p_1^M(\frac{\lambda}{2}) = 0.125214$
3	$p_3 = p_2(\frac{\lambda}{3}) = 0.180447$	$p_3^T = p_2^T(\frac{\lambda}{3}) = 0.208690$	$p_3^M = p_2^M(\frac{\lambda}{3}) = 0.083476$

## 3.6 Mixture Distributions

In this section, you learn how to:

- Define a mixture distribution when the mixing component is based on a finite number of sub-groups
- Compute mixture distribution probabilities from mixing proportions and knowledge of the distribution of each subgroup
- Define a mixture distribution when the mixing component is continuous

In many applications, the underlying population consists of naturally defined sub-groups with some homogeneity within each sub-group. In such cases it is convenient to model the individual sub-groups, and in a ground-up manner

model the whole population. As we shall see below, beyond the aesthetic appeal of the approach, it also extends the range of applications that can be catered to by standard parametric distributions.

Let  $k$  denote the number of defined sub-groups in a population, and let  $F_i$  denote the distribution of an observation drawn from the  $i$ -th subgroup. If we let  $\alpha_i$  denote the proportion of the population in the  $i$ -th subgroup, with  $\sum_{i=1}^k \alpha_i = 1$ , then the distribution of a randomly chosen observation from the population, denoted by  $F$ , is given by

$$F(x) = \sum_{i=1}^k \alpha_i \cdot F_i(x). \quad (3.6)$$

The above expression can be seen as a direct application of the Law of Total Probability. As an example, consider a population of drivers split broadly into two sub-groups, those with at most five years of driving experience and those with more than five years experience. Let  $\alpha$  denote the proportion of drivers with less than 5 years experience, and  $F_{\leq 5}$  and  $F_{>5}$  denote the distribution of the count of claims in a year for a driver in each group, respectively. Then the distribution of claim count of a randomly selected driver is given by

$$\alpha \cdot F_{\leq 5}(x) + (1 - \alpha)F_{>5}(x).$$

An alternate definition of a mixture distribution is as follows. Let  $N_i$  be a random variable with distribution  $F_i$ ,  $i = 1, \dots, k$ . Let  $I$  be a random variable taking values  $1, 2, \dots, k$  with probabilities  $\alpha_1, \dots, \alpha_k$ , respectively. Then the random variable  $N_I$  has a distribution given by equation (3.6)<sup>7</sup>.

In (3.6) we see that the distribution function is a convex combination of the component distribution functions. This result easily extends to the probability mass function, the survival function, the raw moments, and the expectation as these are all linear mappings of the distribution function. We note that this is not true for central moments like the variance, and conditional measures like the hazard rate function. In the case of variance it is easily seen as

$$\text{Var}[N_I] = \text{E}[\text{Var}[N_I|I]] + \text{Var}[\text{E}[N_I|I]] = \sum_{i=1}^k \alpha_i \text{Var}[N_i] + \text{Var}[\text{E}[N_I|I]]. \quad (3.7)$$

Appendix Chapter ?? provides additional background about this important expression.

---

<sup>7</sup>This in particular lays out a way to simulate from a mixture distribution that makes use of efficient simulation schemes that may exist for the component distributions.

**Example 3.6.1. Actuarial Exam Question.** In a certain town the number of common colds an individual will get in a year follows a Poisson distribution that depends on the individual's age and smoking status. The distribution of the population and the mean number of colds are as follows:

Table 3.2. The Distribution of the Population and the Mean Number of Colds

	Proportion of population	Mean number of colds
Children	0.3	3
Adult Non-Smokers	0.6	1
Adult Smokers	0.1	4

1. Calculate the probability that a randomly drawn person has 3 common colds in a year.
2. Calculate the conditional probability that a person with exactly 3 common colds in a year is an adult smoker.

**Example Solution.**

1. Using Law of Total Probability, we can write the required probability as  $\Pr(N_I = 3)$ , with  $I$  denoting the group of the randomly selected individual with 1, 2 and 3 signifying the groups \*Children\*, \*Adult Non-Smoker\*, and \*Adult Smoker\*, respectively. Now by conditioning we get

$$\Pr(N_I = 3) = 0.3 \cdot \Pr(N_1 = 3) + 0.6 \cdot \Pr(N_2 = 3) + 0.1 \cdot \Pr(N_3 = 3),$$

with  $N_1, N_2$  and  $N_3$  following Poisson distributions with means 3, 1, and 4, respectively. Using the above, we get  $\Pr(N_I = 3) \sim 0.1235$ . The conditional probability of event A given event B,  $\Pr(A|B) = \frac{\Pr(A,B)}{\Pr(B)}$ . The required conditional probability in this problem can then be written as  $\Pr(I = 3|N_I = 3)$ , which equals

$$\Pr(I = 3|N_I = 3) = \frac{\Pr(I = 3, N_3 = 3)}{\Pr(N_I = 3)} \sim \frac{0.1 \times 0.1954}{0.1235} \sim 0.1581.$$

---

In the above example, the number of subgroups  $k$  was equal to three. In general,  $k$  can be any natural number, but when  $k$  is large it is parsimonious from a modeling point of view to take the following *infinitely many subgroup* approach. To motivate this approach, let the  $i$ -th subgroup be such that its component distribution  $F_i$  is given by  $G_{\tilde{\theta}_i}$ , where  $G_{\cdot}$  is a parametric family of distributions with parameter space  $\Theta \subseteq \mathbb{R}^d$ . With this assumption, the

distribution function  $F$  of a randomly drawn observation from the population is given by

$$F(x) = \sum_{i=1}^k \alpha_i G_{\tilde{\theta}_i}(x), \quad \forall x \in \mathbb{R},$$

similar to equation (3.6). Alternately, it can be written as

$$F(x) = \mathbb{E}[G_{\tilde{\vartheta}}(x)], \quad \forall x \in \mathbb{R},$$

where  $\tilde{\vartheta}$  takes values  $\tilde{\theta}_i$  with probability  $\alpha_i$ , for  $i = 1, \dots, k$ . The above makes it clear that when  $k$  is large, one could model the above by treating  $\tilde{\vartheta}$  as continuous random variable.

To illustrate this approach, suppose we have a population of drivers with the distribution of claims for an individual driver being distributed as a Poisson. Each person has their own (personal) expected number of claims  $\lambda$  - smaller values for good drivers, and larger values for others. There is a distribution of  $\lambda$  in the population; a popular and convenient choice for modeling this distribution is a gamma distribution with parameters  $(\alpha, \theta)$  (the gamma distribution will be introduced formally in Section 4.2.1). With these specifications it turns out that the resulting distribution of  $N$ , the claims of a randomly chosen driver, is a negative binomial with parameters  $(r = \alpha, \beta = \theta)$ . This can be shown in many ways, but a straightforward argument is as follows:

$$\begin{aligned} \Pr(N = k) &= \int_0^\infty \frac{e^{-\lambda} \lambda^k}{k!} \frac{\lambda^{\alpha-1} e^{-\lambda/\theta}}{\Gamma(\alpha)\theta^\alpha} d\lambda = \frac{1}{k! \Gamma(\alpha)\theta^\alpha} \int_0^\infty \lambda^{\alpha+k-1} e^{-\lambda(1+1/\theta)} d\lambda \\ &= \frac{\Gamma(\alpha+k)}{k! \Gamma(\alpha)\theta^\alpha (1+1/\theta)^{\alpha+k}} \\ &= \binom{\alpha+k-1}{k} \left(\frac{1}{1+\theta}\right)^\alpha \left(\frac{\theta}{1+\theta}\right)^k, \quad k = 0, 1, \dots \end{aligned}$$

Note that the above derivation implicitly uses the following:

$$f_{N|\Lambda=\lambda}(N = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k \geq 0; \quad \text{and} \quad f_\Lambda(\lambda) = \frac{\lambda^{\alpha-1} e^{-\lambda/\theta}}{\Gamma(\alpha)\theta^\alpha}, \quad \lambda > 0.$$

By considering mixtures of a parametric class of distributions, we increase the richness of the class. This expansion of distributions results in the mixture class being able to cater well to more applications than the parametric class we started with. Mixture modeling is an important modeling technique in insurance applications and later chapters will cover more aspects of this modeling technique.

**Example 3.6.2.** Suppose that  $N|\Lambda \sim \text{Poisson}(\Lambda)$  and that  $\Lambda \sim \text{gamma}$  with mean of 1 and variance of 2. Determine the probability that  $N = 1$ .

**Example Solution.** For a gamma distribution with parameters  $(\alpha, \theta)$ , we have that the mean is  $\alpha\theta$  and the variance is  $\alpha\theta^2$ . Using these expressions we have

$$\alpha = \frac{1}{2} \text{ and } \theta = 2.$$

Now, one can directly use the above result to conclude that  $N$  is distributed as a negative binomial with  $r = \alpha = \frac{1}{2}$  and  $\beta = \theta = 2$ . Thus

$$\begin{aligned}\Pr(N = 1) &= \binom{1+r-1}{1} \left(\frac{1}{(1+\beta)^r}\right) \left(\frac{\beta}{1+\beta}\right)^1 \\ &= \binom{1+\frac{1}{2}-1}{1} \frac{1}{(1+2)^{1/2}} \left(\frac{2}{1+2}\right)^1 \\ &= \frac{1}{3^{3/2}} = 0.19245.\end{aligned}$$

### 3.7 Real Data Example

In this section, you learn how to:

- Compare a fitted distribution to empirical data to assess the adequacy of the fit

In the above, we have discussed three basic frequency distributions, along with their extensions through zero modification/truncation, and by looking at mixtures of these distributions. Nevertheless, these classes remain parametric and hence a small subset of the class of all possible frequency distributions (that is, the set of distributions on non-negative integers). Hence, even though we have discussed methods for estimating the unknown parameters, the *fitted* distribution need not be a good representation of the underlying distribution if the latter is **far** from the class of distribution used for modeling.

While the class of distributions considered above is relatively narrow, via a real example, we present some evidence that they serve insurance purposes quite well.

In 1993, a portfolio of  $n = 7,483$  automobile insurance policies from a major

Singaporean insurance company had the distribution of auto accidents per policyholder as given in [Table 3.3](#).

**Table 3.3. Singaporean Automobile Accident Data**

Count ( $k$ )	0	1	2	3	4	Total
No. of Policies with $k$ accidents ( $m_k$ )	6,996	455	28	4	0	7,483

If we fit a Poisson distribution, then the *mle* for  $\lambda$ , the Poisson mean, is the sample mean which is given by

$$\bar{N} = \frac{0 \cdot 6996 + 1 \cdot 455 + 2 \cdot 28 + 3 \cdot 4 + 4 \cdot 0}{7483} = 0.06989.$$

Now if we use Poisson ( $\hat{\lambda}_{MLE}$ ) as the fitted distribution, then a tabular comparison of the fitted counts and observed counts is given by [Table 3.4](#) below, where  $\hat{p}_k$  represents the estimated probabilities under the fitted Poisson distribution.

**Table 3.4. Comparison of Observed to Fitted Counts: Singaporean Auto Data**

Count ( $k$ )	Observed ( $m_k$ )	Fitted Counts Using Poisson ( $n\hat{p}_k$ )
0	6,996	6,977.86
1	455	487.70
2	28	17.04
3	4	0.40
$\geq 4$	0	0.01
Total	7,483	7,483.00

Notice that the fit seems *quite reasonable* from the above tabular comparison, suggesting that the Poisson distribution is a good model of the underlying distribution. Nevertheless, it is worth pointing out that such a tabular comparison falls short of a statistical test of the hypothesis that the underlying distribution is indeed Poisson. In Section ??, we present Pearson's chi-square statistic as a goodness-of-fit statistical measure for this purpose.

### 3.8 Exercises

#### Theoretical Exercises

**Exercise 3.1.** Derive an expression for  $p_N(\cdot)$  in terms of  $F_N(\cdot)$  and  $S_N(\cdot)$ .

**Exercise 3.2.** A measure of center of location must be **equi-variant** with respect to shifts, or location transformations. In other words, if  $N_1$  and  $N_2$  are two random variables such that  $N_1 + c$  has the same distribution as  $N_2$ , for some constant  $c$ , then the difference between the measures of the center of location of  $N_2$  and  $N_1$  must equal  $c$ . Show that the mean satisfies this property.

**Exercise 3.3.** Measures of dispersion should be invariant with respect to shifts and scale equi-variant. Show that standard deviation satisfies these properties by doing the following:

- Show that for a random variable  $N$ , its standard deviation equals that of  $N + c$ , for any constant  $c$ .
- Show that for a random variable  $N$ , its standard deviation equals  $1/c$  times that of  $cN$ , for any positive constant  $c$ .

**Exercise 3.4.** Let  $N$  be a random variable with probability mass function given by

$$p_N(k) = \begin{cases} \left(\frac{6}{\pi^2}\right)\left(\frac{1}{k^2}\right), & k \geq 1; \\ 0, & \text{otherwise.} \end{cases}$$

Show that the mean of  $N$  is  $\infty$ .

**Exercise 3.5.** Let  $N$  be a random variable with a finite second moment. Show that the function  $\psi(\cdot)$  defined by

$$\psi(x) = E(N - x)^2. \quad x \in \mathbb{R}$$

is minimized at  $\mu_N$  without using calculus. Also, give a proof of this fact using derivatives. Conclude that the minimum value equals the variance of  $N$ .

**Exercise 3.6.** Derive the first two central moments of the  $(a, b, 0)$  distributions using the methods mentioned below:

- For the binomial distribution, derive the moments using only its *pmf*, then its *mgf*, and then its *pgf*.
- For the Poisson distribution, derive the moments using only its *mgf*.
- For the negative binomial distribution, derive the moments using only its *pmf*, and then its *pgf*.

**Exercise 3.7.** Let  $N_1$  and  $N_2$  be two independent Poisson random variables with means  $\lambda_1$  and  $\lambda_2$ , respectively. Identify the conditional distribution of  $N_1$  given  $N_1 + N_2$ .

**Exercise 3.8. (Non-Uniqueness of the MLE)** Consider the following parametric family of densities indexed by the parameter  $p$  taking values in  $[0, 1]$ :

$$f_p(x) = p \cdot \phi(x+2) + (1-p) \cdot \phi(x-2), \quad x \in \mathbb{R},$$

where  $\phi(\cdot)$  represents the standard normal density.

- Show that for all  $p \in [0, 1]$ ,  $f_p(\cdot)$  above is a valid density function.
- Find an expression in  $p$  for the mean and the variance of  $f_p(\cdot)$ .
- Let us consider a sample of size one consisting of  $x$ . Show that when  $x$  equals 0, the set of *maximum likelihood estimates* for  $p$  equals  $[0, 1]$ ; also show that the *mle* is unique otherwise.

**Exercise 3.9.** Graph the region of the plane corresponding to values of  $(a, b)$  that give rise to valid  $(a, b, 0)$  distributions. Do the same for  $(a, b, 1)$  distributions.

**Exercise 3.10. (Computational Complexity)** For the  $(a, b, 0)$  class of distributions, count the number of basic mathematical operations (addition, subtraction, multiplication, division) needed to compute the  $n$  probabilities  $p_0 \dots p_{n-1}$  using the recurrence relationship. For the negative binomial distribution with non-integer  $r$ , count the number of such operations. What do you observe?

**Exercise 3.11.** Using the development of Section 3.3 rigorously show that not only does the recurrence (3.1) tie the binomial, the Poisson and the negative binomial distributions together, but that it also characterizes them.

**Exercise 3.12. Actuarial Exam Question.** You are given:

1.  $p_k$  denotes the probability that the number of claims equals  $k$  for  $k = 0, 1, 2, \dots$
2.  $\frac{p_n}{p_m} = \frac{m!}{n!}, m \geq 0, n \geq 0$

Using the corresponding zero-modified claim count distribution with  $p_0^M = 0.1$ , calculate  $p_1^M$ .

#### Exercises with a Practical Focus

**Exercise 3.13. Singaporean Automobile Accident.** In this exercise, we replicate and extend the real-data example introduced in Section 3.7 using R.

- a. From the package **CASdatasets**, retrieve the data **sgautonb** in order to work with the variable **C1m\_Count** which is a count of claims. Refer to Section ?? for a description of this package. Verify that the mean claim count is 0.
- b. Compute the fitted Poisson distribution and reproduce Table 3.5.
- c. Compute the maximum likelihood estimates for the negative binomial distribution. One way to do this is to create a negative logarithmic likelihood function and use the R function **optim** for minimization. Use the resulting

**TABLE 3.5: Singaporean Automobile Comparison of Empirical to Poisson Fitted Percentiles**

Claim Count	Empirical Percentile	Poisson Perc
6996	0.9349	0.9325
455	0.9957	0.9977
28	0.9995	0.9999
4	1.0000	1.0000

maximum likelihood estimates to create a fitted distribution and augment the Table in part (b) with this alternative distribution.

In part (c), you learn that the more complex negative binomial distribution produces roughly the same fits as the simpler Poisson distribution. As a result of this analysis, an analyst would typically prefer the simpler Poisson distribution.

**Exercise 3.14. Corporate Travel.** This exercise is based on the data set introduced in [Exercise 1.1](#) where now the focus is on frequency modeling. For corporate travel, the number of claims are sufficient that a separate frequency model could be considered. For the frequency of claims, there are 2107 claims over the 2006-2021 period that amounts to 131.69 per year. One might assume that annual claims can be fit using a single distribution to the entire period, such as a Poisson or a negative binomial. Another option is to fit a distribution starting in years 2009, where this is an increase in the amount of claims from prior years. A third option is to omit experience from underwriting year 2019 and on where the number of claims fluctuated dramatically, in part due to the Covid epidemic. In this exercise, we pursue the first option.

- a. Fit a Poisson distribution and a negative binomial distribution to all claims.
- b. Fit a negative binomial distribution to all claims using the strategy introduced in part (c) of Exercise 3.13.
- c. To check your work, use the `fitdist` function from the package `fitdistrplus`, with the negative binomial (`nbinom`) option.
- d. Use the `ecdf` function in R to produce empirical cumulative probabilities. Produce a table that compares the empirical percentiles to those under the Poisson and negative binomial.

From part (d), you learn that both fitted distributions did well and neither outperformed the other.

### 3.9 Further Resources and Contributors

Appendix Chapter ?? gives a general introduction to maximum likelihood theory regarding estimation of parameters from a parametric family. Appendix Chapter ?? gives more specific examples and expands some of the concepts.

If you would like additional practice with R coding, please visit our companion [LDA Short Course](#). In particular, see the [Frequency Modeling Chapter](#).

#### Contributors

- **N.D. Shyamalkumar**, The University of Iowa, and **Krupa Viswanathan**, Temple University, are the principal authors of the initial version and also the second edition of this chapter. Email: [shyamal-kumar@uiowa.edu](mailto:shyamal-kumar@uiowa.edu) for chapter comments and suggested improvements.
- Chapter reviewers include: Chunsheng Ban, Paul Johnson, Hirokazu (Iwahiro) Iwasawa, Dalia Khalil, Tatjana Miljkovic, Rajesh Sahasrabuddhe, and Michelle Xia.

### 3.9.1 TS 3.A. R Code for Plots

**Code for Figure 3.2:**

```

likm<-function(m){
  prod((dbinom(x,m,mean(x)/m)))
}
x<-c(2,2,2,4,5);
n<-(5:100);
# Computing the Likelihood
ll<-sapply(n,likm);
# Computing the MLE
n[ll==max(ll)];
# Storing the Likelihood Curve
y<-cbind(n,ll);

# Second Dataset
x<-c(2,2,2,4,6);
ll<-sapply(n,likm);
n[ll==max(ll)];
y<-cbind(y,ll);

# Third Dataset
x<-c(2,2,2,4,7);
ll<-sapply(n,likm);
n[ll==max(ll)];
y<-cbind(y,ll);

colnames(y)<-c("m","$\\tilde{x}=(2,2,2,4,5)$",
               "$\\tilde{x}=(2,2,2,4,6)$",
               "$\\tilde{x}=(2,2,2,4,7)$");
dy<-data.frame(y);
library(tikzDevice);
library(ggplot2);
options(tikzMetricPackages =
        c("\\usepackage[utf8]{inputenc}", "\\usepackage[T1]{fontenc}",
          "\\usetikzlibrary{calc}", "\\usepackage{amssymb}",
          "\\usepackage{amsmath}", "\\usepackage[active]{preview}"))
tikz(file = "plot_test.tex", width = 6.25, height = 3.125);
ggplot(dy) +
  geom_point(aes(x=m, y=(X..tilde.x...2.2.2.4.5..),
                 shape="$\\tilde{x}=(2,2,2,4,5):\\hat{m}=7$"), size=0.75) +
  geom_point(aes(x=m, y=(X..tilde.x...2.2.2.4.6..),
                 shape="$\\tilde{x}=(2,2,2,4,6):\\hat{m}=18$"), size=0.75) +
  geom_point(aes(x=m, y=(X..tilde.x...2.2.2.4.7..),
                 shape="$\\tilde{x}=(2,2,2,4,7):\\hat{m}=\\infty$"), size=0.75) +
  geom_point(aes(x=c(7),y=dy$X..tilde.x...2.2.2.4.5..[3],colour="$\\hat{m}$",
                 shape="$\\tilde{x}=(2,2,2,4,5):\\hat{m}=7$"), size=0.75) +
  geom_point(aes(x=c(18),y=dy$X..tilde.x...2.2.2.4.6..[14],colour="$\\hat{m}$",
                 shape="$\\tilde{x}=(2,2,2,4,6):\\hat{m}=18$"), size=0.75) +
  labs(x="m",y="$L(m,\\overline{x}/m)$",
       title="MLE for $m$: Non-Robustness of MLE ");
dev.off();

```

### Code for Figure 3.3:

```

likbinm<-function(m){
  # binomial likelihood maximized w.r.t. p
  prod((dbinom(x,m,mean(x)/m)))
}

liknbinm<-function(r){
  # negative binomial likelihood maximized w.r.t. beta
  prod(dnbinom(x,r,1-mean(x)/(mean(x)+r)))
}

# Data Matrix; Three Samples, one in each Column;
# First Sample has Var<Mean
# Second Sample has Var=Mean
# Third Sample has Var>Mean

X<-cbind(c(2,5,6,8,9)+2,c(2,5,6,8,9),c(2,3,6,8,9))

# Used for creating the labels in the z matrix
ord_char<-c("<","=",>")

# Empty matrices;
Y<-matrix(1,ncol=2,nrow=0)
Z<-matrix(1,ncol=2,nrow=0)

for (i in (1:3)) {
  # Work with data in the i-th sample
  x<-X[,i]

  # Binomial Likelihood
  # Interval of n values covering the MLE
  n<-(9:100)
  # Evaluating the Likelihood at various values of n
  ll<-sapply(n,likbinm)
  # Finding the MLE of n
  n[ll==max(ll[!is.na(ll)])]
  # Storing the data and the labels
  Y<-rbind(Y,cbind(n,ll))
  Z<-rbind(Z,cbind(rep(paste("$\\hat{\\sigma}^2",ord_char[i],
    "\\hat{\\mu}$"),length(n)),rep("Binomial - L(m,\\overline{x}/m)$",
    length(n)))))

  # Negative Binomial Likelihood
  # Interval of r values
  r<-(1:100)
  # Evaluating the Likelihood at various values of r
  ll<-sapply(r,liknbinm)
  # Finding the MLE of r
  ll[is.na(ll)]=0
  r[ll==max(ll[!is.na(ll)])]
  # Storing the data and the labels
  Y<-rbind(Y,cbind(r,ll))
  Z<-rbind(Z,cbind(rep(paste("$\\hat{\\sigma}^2",ord_char[i],
    "\\hat{\\mu}$"),length(r)),rep("Neg.Binomial -$L(r,\\overline{x}/r)$",
    length(r))))
}

```

```
length(r))))  
  
# Poisson Likelihood  
# Storing the data and the labels  
# In the Poisson case MLE is the sample mean  
Y<-rbind(Y,cbind(r,rep(prod(dpois(x,mean(x))),length(r))))  
Z<-rbind(Z,cbind(rep(paste("$\\hat{\\sigma}^2",ord_char[i],"\\hat{\\mu}"),  
length(r)),rep("Poisson - $L(\\overline{x})$",length(r))))  
}  
  
# Assigning Column Names  
colnames(Y)<-c("x","lik")  
colnames(Z)<-c("dataset","Distribution")  
# Creating a Dataframe for using ggplot  
dy<-cbind(data.frame(Y),data.frame(Z))  
  
library(tikzDevice)  
library(ggplot2)  
options(tikzMetricPackages = c("\\usepackage[utf8]{inputenc}",  
"\\usepackage[T1]{fontenc}",  
"\\usetikzlibrary{calc}",  
"\\usepackage{amssymb}",  
"\\usepackage{amsmath}",  
"\\usepackage[active]{preview}" ))  
tikz(file = "plot_test_2.tex", width = 6.25, height = 6.25)  
ggplot(data=dy,aes(x=x,y=lik,col=Distribution)) +  
  geom_point(size=0.25) + facet_grid(dataset~.) +  
  labs(x="m/r",y="Likelihood",title="")  
dev.off()
```



# 4

---

## *Modeling Loss Severity*

---

*Chapter Preview.* The traditional loss distribution approach to modeling aggregate losses starts by separately fitting a frequency distribution to the number of losses and a severity distribution to the size of losses. The estimated aggregate loss distribution combines the loss frequency distribution and the loss severity distribution by convolution. Discrete distributions often referred to as counting or frequency distributions were used in Chapter 3 to describe the number of events such as number of accidents to the driver or number of claims to the insurer. Lifetimes, asset values, losses and claim sizes are usually modeled as continuous random variables and as such are modeled using continuous distributions, often referred to as loss or severity distributions. A mixture distribution is a weighted combination of simpler distributions that is used to model phenomenon investigated in a heterogeneous population, such as modeling more than one type of claims in liability insurance (small frequent claims and large relatively rare claims). In this chapter we explore the use of continuous as well as mixture distributions to model the random size of loss.

Sections 4.1 to 4.3 present key attributes that characterize continuous models and means of creating new distributions from existing ones. Section 4.4.1 describes some principal non-parametric methods for estimating loss distributions: moment and percentile based, empirical, and density estimation methods. Section 4.4.2 covers parametric estimation methods including method of moments and percentile matching, and deepens our understanding of maximum likelihood methods. The frequency distributions from Chapter 3 will be combined with the ideas from this chapter to describe the aggregate losses over the whole portfolio in Chapter 7.

---

### **4.1 Basic Distributional Quantities**

---

In this section, you learn how to define some basic distributional quantities:

- moments,

- moment generating functions, and
  - percentiles
- 

#### 4.1.1 Moments and Moment Generating Functions

Let  $X$  be a continuous random variable with probability density function (*pdf*)  $f_X(x)$  and distribution function  $F_X(x)$ . The  $k$ -th raw moment of  $X$ , denoted by  $\mu'_k$ , is the expected value of the  $k$ -th power of  $X$ , provided it exists. The first raw moment  $\mu'_1$  is the mean of  $X$  usually denoted by  $\mu$ . The formula for  $\mu'_k$  is given as

$$\mu'_k = E(X^k) = \int_0^\infty x^k f_X(x) dx.$$

Note that the notation used here for moments differs from the notation used in Section 3.2.1. The support of the random variable  $X$  is assumed to be non-negative since actuarial phenomena are rarely negative. For example, an easy integration by parts shows that the raw moments for nonnegative variables can also be computed using

$$\mu'_k = \int_0^\infty k x^{k-1} [1 - F_X(x)] dx,$$

that is based on the survival function, denoted as  $S_X(x) = 1 - F_X(x)$ . This formula is particularly useful when  $k = 1$ . Section 5.1.2 discusses this approach in more detail.

The  $k$ -th central moment of  $X$ , denoted by  $\mu_k$ , is the expected value of the  $k$ -th power of the deviation of  $X$  from its mean  $\mu$ . The formula for  $\mu_k$  is given as

$$\mu_k = E[(X - \mu)^k] = \int_0^\infty (x - \mu)^k f_X(x) dx.$$

The second central moment  $\mu_2$  defines the variance of  $X$ , denoted by  $\sigma^2$ . The square root of the variance is the standard deviation  $\sigma$ .

From a classical perspective, further characterization of the shape of the distribution includes its degree of symmetry as well as its flatness compared to the normal distribution. The ratio of the third central moment to the cube of the standard deviation ( $\mu_3/\sigma^3$ ) defines the coefficient of skewness which is a measure of symmetry. A positive coefficient of skewness indicates that the distribution is skewed to the right (positively skewed). The ratio of the fourth central moment to the fourth power of the standard deviation ( $\mu_4/\sigma^4$ ) defines the coefficient of kurtosis. The normal distribution has a coefficient of kurtosis of 3. Distributions with a coefficient of kurtosis greater than 3 have

heavier tails than the normal, whereas distributions with a coefficient of kurtosis less than 3 have lighter tails and are flatter. Section ?? describes the tails of distributions from an insurance and actuarial perspective.

**Example 4.1.1. Actuarial Exam Question.** Assume that the rv  $X$  has a gamma distribution with mean 8 and skewness 1. Find the variance of  $X$ . (*Hint:* The gamma distribution is reviewed in Section 4.2.1.)

**Example Solution.** The *pdf* of  $X$  is given by

$$f_X(x) = \frac{(x/\theta)^\alpha}{x \Gamma(\alpha)} e^{-x/\theta}$$

for  $x > 0$ . For  $\alpha > 0$ , the  $k$ -th raw moment is

$$\mu'_k = E(X^k) = \int_0^\infty \frac{1}{\Gamma(\alpha) \theta^\alpha} x^{k+\alpha-1} e^{-x/\theta} dx = \frac{\Gamma(k+\alpha)}{\Gamma(\alpha)} \theta^k$$

Given  $\Gamma(r+1) = r\Gamma(r)$  and  $\Gamma(1) = 1$ , then  $\mu'_1 = E(X) = \alpha\theta$ ,  $\mu'_2 = E(X^2) = (\alpha+1)\alpha\theta^2$ ,  $\mu'_3 = E(X^3) = (\alpha+2)(\alpha+1)\alpha\theta^3$ , and  $\text{Var}(X) = (\alpha+1)\alpha\theta^2 - (\alpha\theta)^2 = \alpha\theta^2$ .

$$\begin{aligned} \text{Skewness} &= \frac{E[(X-\mu'_1)^3]}{(\text{Var } X)^{3/2}} = \frac{\mu'_3 - 3\mu'_2\mu'_1 + 2\mu'_1^3}{(\text{Var } X)^{3/2}} \\ &= \frac{(\alpha+2)(\alpha+1)\alpha\theta^3 - 3(\alpha+1)\alpha^2\theta^3 + 2\alpha^3\theta^3}{(\alpha\theta^2)^{3/2}} \\ &= \frac{2}{\alpha^{1/2}} = 1. \end{aligned}$$

Hence,  $\alpha = 4$ . Since,  $E(X) = \alpha\theta = 8$ , then  $\theta = 2$  and finally,  $\text{Var}(X) = \alpha\theta^2 = 16$ .

The moment generating function (mgf), denoted by  $M_X(t)$  uniquely characterizes the distribution of  $X$ . While it is possible for two different distributions to have the same moments and yet still differ, this is not the case with the moment generating function. That is, if two random variables have the same moment generating function, then they have the same distribution. The moment generating function is given by

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} f_X(x) dx$$

for all  $t$  for which the expected value exists. The *mgf* is a real function whose  $k$ -th derivative at zero is equal to the  $k$ -th raw moment of  $X$ . In symbols, this

is

$$\frac{d^k}{dt^k} M_X(t) \Big|_{t=0} = E(X^k).$$

**Example 4.1.2. Actuarial Exam Question.** The random variable  $X$  has an exponential distribution with mean  $\frac{1}{b}$ . It is found that  $M_X(-b^2) = 0.2$ . Find  $b$ . (*Hint:* The exponential is a special case of the gamma distribution which is reviewed in Section 4.2.1.)

**Example Solution.** With  $X$  having an exponential distribution with mean  $\frac{1}{b}$ , we have that

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} b e^{-bx} dx = \int_0^\infty b e^{-x(b-t)} dx = \frac{b}{(b-t)}.$$

Then,

$$M_X(-b^2) = \frac{b}{(b+b^2)} = \frac{1}{(1+b)} = 0.2.$$

Thus,  $b = 4$ .

**Example 4.1.3. Actuarial Exam Question.** Let  $X_1, \dots, X_n$  be independent random variables, where  $X_i$  has a gamma distribution with parameters  $\alpha_i$  and  $\theta$ . Find the distribution of  $S = \sum_{i=1}^n X_i$ , the mean  $E(S)$ , and the variance  $\text{Var}(S)$ .

**Example Solution.** The mgf of  $S$  is

$$M_S(t) = E(e^{tS}) = E\left(e^{t\sum_{i=1}^n X_i}\right) = E\left(\prod_{i=1}^n e^{tX_i}\right).$$

Using independence, we get

$$M_S(t) = \prod_{i=1}^n E(e^{tX_i}) = \prod_{i=1}^n M_{X_i}(t).$$

The moment generating function of the gamma distribution  $X_i$  is  $M_{X_i}(t) = (1-\theta t)^{-\alpha_i}$ . Then,

$$M_S(t) = \prod_{i=1}^n (1-\theta t)^{-\alpha_i} = (1-\theta t)^{-\sum_{i=1}^n \alpha_i}.$$

This indicates that the distribution of  $S$  is gamma with parameters  $\sum_{i=1}^n \alpha_i$  and  $\theta$ .

This is a demonstration of how we can use the uniqueness property of the moment generating function to determine the probability distribution of a function of random variables.

We can find the mean and variance from the properties of the gamma distribution. Alternatively, by finding the first and second derivatives of  $M_S(t)$  at zero, we can show that  $E(S) = \frac{\partial M_S(t)}{\partial t} \Big|_{t=0} = \alpha\theta$  where  $\alpha = \sum_{i=1}^n \alpha_i$ , and

$$E(S^2) = \left. \frac{\partial^2 M_S(t)}{\partial t^2} \right|_{t=0} = (\alpha + 1)\alpha\theta^2.$$

Hence,  $\text{Var}(S) = \alpha\theta^2$ .

---

One can also use the moment generating function to compute the probability generating function

$$P_X(z) = E(z^X) = M_X(\log z).$$

As introduced in Section 3.2.2, the probability generating function is more useful for discrete random variables.

---

### 4.1.2 Quantiles

Quantiles can also be used to describe the characteristics of the distribution of  $X$ . When the distribution of  $X$  is continuous, for a given fraction  $0 \leq p \leq 1$  the corresponding quantile is the solution of the equation

$$F_X(\pi_p) = p.$$

For example, the middle point of the distribution,  $\pi_{0.5}$ , is the median. A percentile is a type of quantile; a  $100p$  percentile is the number such that  $100 \times p$  percent of the data is below it.

**Example 4.1.4. Actuarial Exam Question.** Let  $X$  be a continuous random variable with density function  $f_X(x) = \theta e^{-\theta x}$ , for  $x > 0$  and 0 elsewhere. If the median of this distribution is  $\frac{1}{3}$ , find  $\theta$ .

**Example Solution.** The distribution function is  $F_X(x) = 1 - e^{-\theta x}$ . So,

$F_X(\pi_{0.5}) = 1 - e^{-\theta\pi_{0.5}} = 0.5$ . As,  $\pi_{0.5} = \frac{1}{3}$ , we have  $F_X\left(\frac{1}{3}\right) = 1 - e^{-\theta/3} = 0.5$  and  $\theta = 3 \log 2$ .

---

Section 4.4.1 extends the definition of quantiles to include distributions that are discrete, continuous, or a hybrid combination.

---



---

## 4.2 Continuous Distributions for Modeling Loss Severity

---

In this section, you learn how to define and apply four fundamental severity distributions:

- gamma,
  - Pareto,
  - Weibull, and
  - generalized beta distribution of the second kind.
- 

### 4.2.1 Gamma Distribution

Recall that the traditional approach in modeling losses is to fit separate models for frequency and claim severity. When frequency and severity are modeled separately it is common for actuaries to use the Poisson distribution (introduced in Section ??) for claim count and the gamma distribution to model severity. An alternative approach for modeling losses that has recently gained popularity is to create a single model for pure premium (average claim cost).

The continuous variable  $X$  is said to have the gamma distribution with shape parameter  $\alpha$  and scale parameter  $\theta$  if its probability density function is given by

$$f_X(x) = \frac{(x/\theta)^\alpha}{x \Gamma(\alpha)} \exp(-x/\theta) \quad \text{for } x > 0.$$

Note that  $\alpha > 0$ ,  $\theta > 0$ .

The two panels in Figure 4.1 demonstrate the effect of the scale and shape parameters on the gamma density function.

When  $\alpha = 1$  the gamma reduces to an exponential distribution and when  $\alpha = \frac{n}{2}$  and  $\theta = 2$  the gamma reduces to a chi-square distribution with  $n$

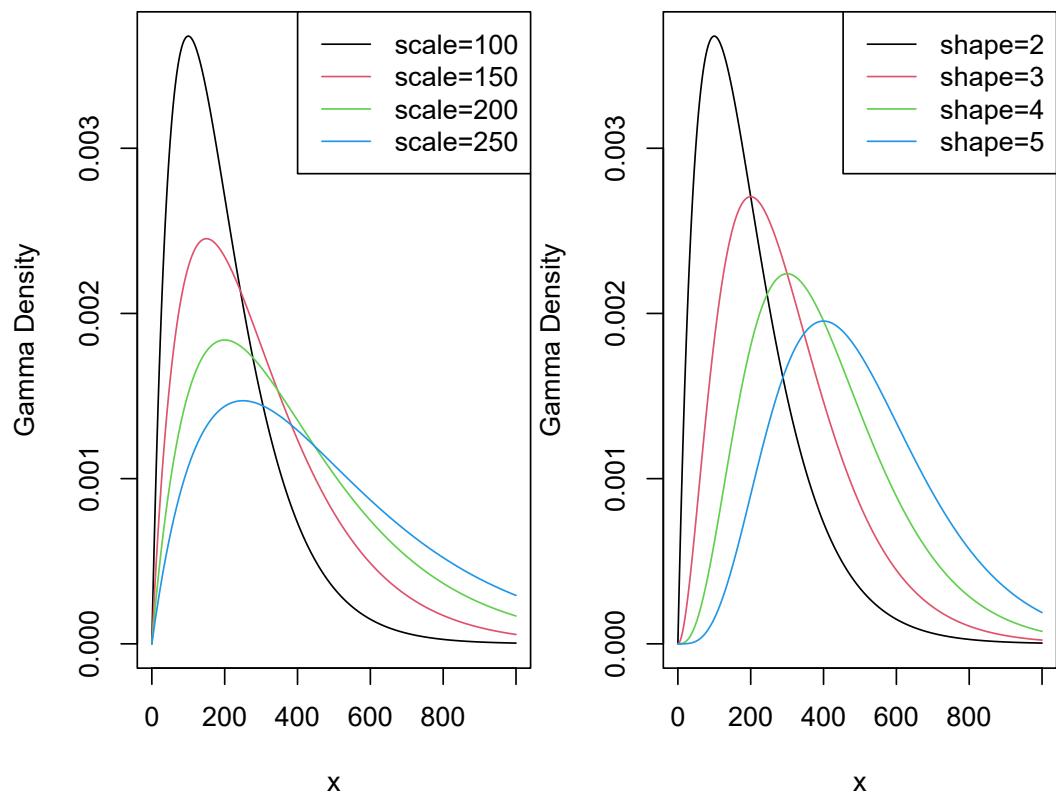


FIGURE 4.1: **Gamma Densities.** The left-hand panel is with shape=2 and varying scale. The right-hand panel is with scale=100 and varying shape.

degrees of freedom. As we will see in Section ??, the chi-square distribution is used extensively in statistical hypothesis testing.

The distribution function of the gamma model is the *incomplete gamma function*, denoted by  $\Gamma(\alpha; \frac{x}{\theta})$ , and defined as

$$F_X(x) = \Gamma\left(\alpha; \frac{x}{\theta}\right) = \frac{1}{\Gamma(\alpha)} \int_0^{x/\theta} t^{\alpha-1} e^{-t} dt,$$

with  $\alpha > 0$ ,  $\theta > 0$ . For an integer  $\alpha$ , it can be written as  $\Gamma(\alpha; \frac{x}{\theta}) = 1 - e^{-x/\theta} \sum_{k=0}^{\alpha-1} \frac{(x/\theta)^k}{k!}$ .

The  $k$ -th raw moment of the gamma distributed random variable for any positive  $k$  is given by

$$\mathbb{E}(X^k) = \theta^k \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)}.$$

The mean and variance are given by  $\mathbb{E}(X) = \alpha\theta$  and  $\text{Var}(X) = \alpha\theta^2$ , respectively.

Since all moments exist for any positive  $k$ , the gamma distribution is considered a light tailed distribution, which may not be suitable for modeling risky assets as it will not provide a realistic assessment of the likelihood of severe losses.

#### 4.2.2 Pareto Distribution

The Pareto distribution, named after the Italian economist Vilfredo Pareto (1843-1923), has many economic and financial applications. It is a positively skewed and heavy-tailed distribution which makes it suitable for modeling income, high-risk insurance claims and severity of large casualty losses. The survival function of the Pareto distribution which decays slowly to zero was first used to describe the distribution of income where a small percentage of the population holds a large proportion of the total wealth. For extreme insurance claims, the tail of the severity distribution (losses in excess of a threshold) can be modeled using a Generalized Pareto distribution.

The continuous variable  $X$  is said to have the (two parameter) Pareto distribution with shape parameter  $\alpha$  and scale parameter  $\theta$  if its pdf is given by

$$f_X(x) = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}} \quad x > 0, \alpha > 0, \theta > 0. \quad (4.1)$$

The two panels in Figure 4.2 demonstrate the effect of the scale and shape parameters on the Pareto density function. There are other formulations of the Pareto distribution including a one parameter version given in Appendix

Section ???. Henceforth, when we refer the Pareto distribution, we mean the version given through the *pdf* in equation (4.1).

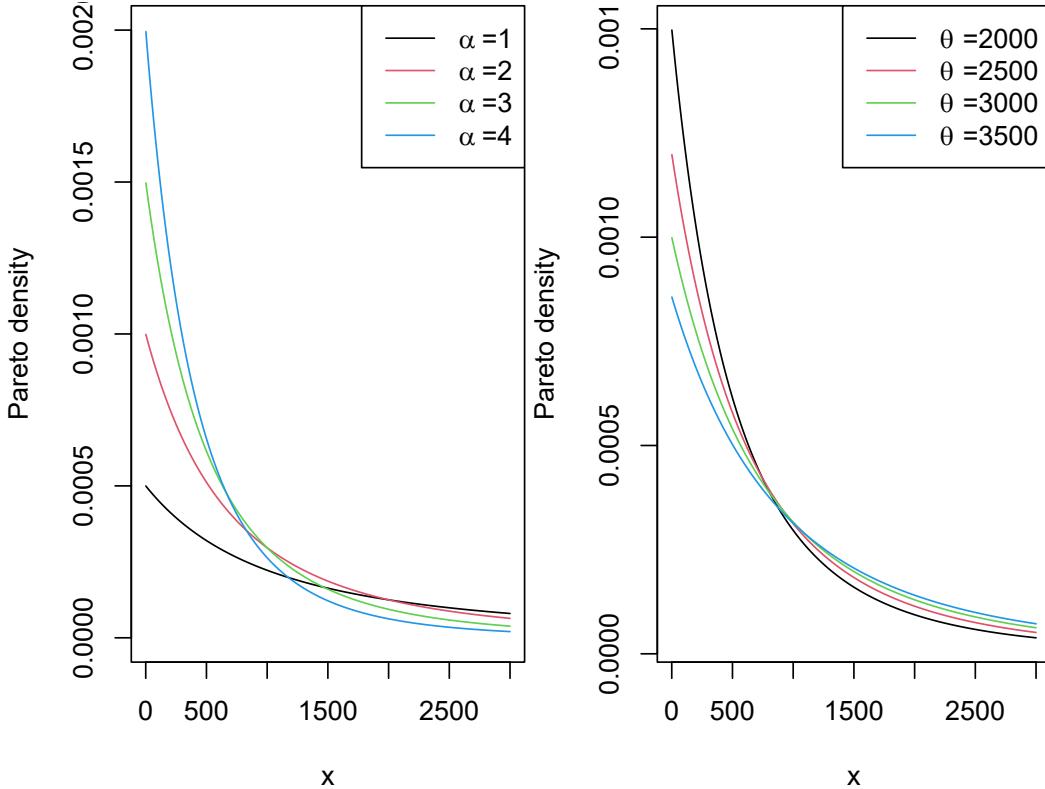


FIGURE 4.2: **Pareto Densities.** The left-hand panel is with scale=2000 and varying shape. The right-hand panel is with shape=3 and varying scale.

The distribution function of the Pareto distribution is given by

$$F_X(x) = 1 - \left( \frac{\theta}{x + \theta} \right)^\alpha \quad x > 0, \alpha > 0, \theta > 0.$$

It can be easily seen that the hazard function of the Pareto distribution is a decreasing function in  $x$ , another indication that the distribution is heavy tailed. Again using the analogy of the income of a population, when the hazard function decreases over time the population dies off at a decreasing rate resulting in a heavier tail for the distribution. The hazard function reveals information about the tail distribution and is often used to model data distributions in survival analysis. The hazard function is defined as the instantaneous potential that the event of interest occurs within a very narrow time frame.

The  $k$ -th raw moment of the Pareto distributed random variable exists, if and

only if,  $\alpha > k$ . If  $k$  is a positive integer then

$$\mathbb{E}(X^k) = \frac{\theta^k k!}{(\alpha-1)\cdots(\alpha-k)} \quad \alpha > k.$$

The mean and variance are given by

$$\mathbb{E}(X) = \frac{\theta}{\alpha-1} \quad \text{for } \alpha > 1$$

and

$$\text{Var}(X) = \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)} \quad \text{for } \alpha > 2,$$

respectively.

**Example 4.2.1.** The claim size of an insurance portfolio follows the Pareto distribution with mean and variance of 40 and 1800, respectively. Find

- a. The shape and scale parameters.
- b. The 95-th percentile of this distribution.

**Example Solution.**

a. As,  $X \sim Pa(\alpha, \theta)$ , we have  $\mathbb{E}(X) = \frac{\theta}{\alpha-1} = 40$  and  $\text{Var}(X) = \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)} = 1800$ . By dividing the square of the first equation by the second we get  $\frac{\alpha-2}{\alpha} = \frac{40^2}{1800}$ . Thus,  $\alpha = 18.02$  and  $\theta = 680.72$ .

b. The 95-th percentile,  $\pi_{0.95}$ , satisfies the equation

$$F_X(\pi_{0.95}) = 1 - \left( \frac{680.72}{\pi_{0.95} + 680.72} \right)^{18.02} = 0.95.$$

Thus,  $\pi_{0.95} = 122.96$ .

### 4.2.3 Weibull Distribution

The Weibull distribution, named after the Swedish physicist Waloddi Weibull (1887-1979) is widely used in reliability, life data analysis, weather forecasts and general insurance claims. Truncated data arise frequently in insurance studies. The Weibull distribution has been used to model excess of loss treaty over automobile insurance as well as earthquake inter-arrival times.

The continuous variable  $X$  is said to have the Weibull distribution with shape

parameter  $\alpha$  and scale parameter  $\theta$  if its *pdf* is given by

$$f_X(x) = \frac{\alpha}{\theta} \left( \frac{x}{\theta} \right)^{\alpha-1} \exp \left( -\left( \frac{x}{\theta} \right)^\alpha \right) \quad x > 0, \alpha > 0, \theta > 0.$$

The two panels in Figure 4.3 demonstrate the effects of the scale and shape parameters on the Weibull density function.

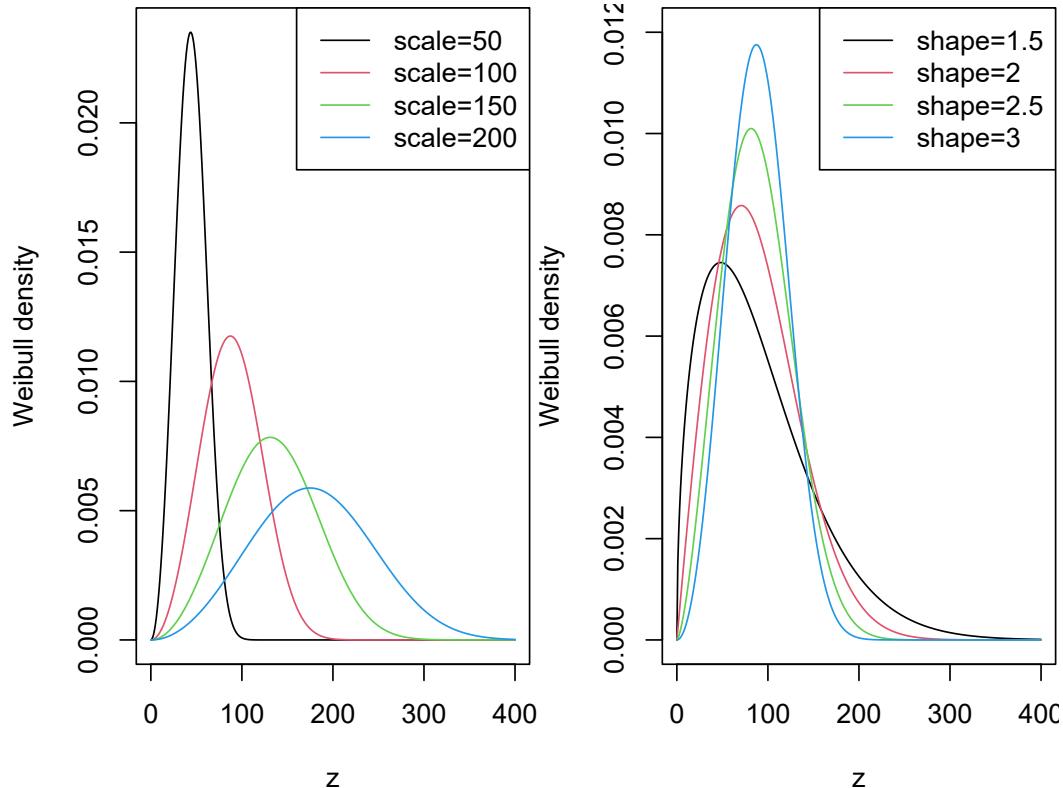


FIGURE 4.3: **Weibull Densities.** The left-hand panel is with shape=3 and varying scale. The right-hand panel is with scale=100 and varying shape.

The distribution function of the Weibull distribution is given by

$$F_X(x) = 1 - \exp \left( -\left( \frac{x}{\theta} \right)^\alpha \right) \quad x > 0, \alpha > 0, \theta > 0.$$

It can be easily seen that the shape parameter  $\alpha$  describes the shape of the hazard function of the Weibull distribution. The hazard function is a decreasing function when  $\alpha < 1$  (heavy tailed distribution), constant when  $\alpha = 1$  and increasing when  $\alpha > 1$  (light tailed distribution). This behavior of the hazard function makes the Weibull distribution a suitable model for a wide variety of phenomena such as weather forecasting, electrical and industrial engineering, insurance modeling, and financial risk analysis.

The  $k$ -th raw moment of the Weibull distributed random variable is given by

$$\mathbb{E}(X^k) = \theta^k \Gamma\left(1 + \frac{k}{\alpha}\right).$$

The mean and variance are given by

$$\mathbb{E}(X) = \theta \Gamma\left(1 + \frac{1}{\alpha}\right)$$

and

$$\text{Var}(X) = \theta^2 \left( \Gamma\left(1 + \frac{2}{\alpha}\right) - \left[ \Gamma\left(1 + \frac{1}{\alpha}\right) \right]^2 \right),$$

respectively.

**Example 4.2.2.** Suppose that the probability distribution of the lifetime of AIDS patients (in months) from the time of diagnosis is described by the Weibull distribution with shape parameter 1.2 and scale parameter 33.33.

- a. Find the probability that a randomly selected person from this population survives at least 12 months.
- b. A random sample of 10 patients will be selected from this population. What is the probability that at most two will die within one year of diagnosis.
- c. Find the 99-th percentile of the distribution of lifetimes.

**Example Solution.**

- a. Let  $X$  be the lifetime of AIDS patients (in months) having a Weibull distribution with parameters (1.2, 33.33). We have,

$$\Pr(X \geq 12) = S_X(12) = e^{-(\frac{12}{33.33})^{1.2}} = 0.746.$$

- b. Let  $Y$  be the number of patients who die within one year of diagnosis. Then,  $Y \sim \text{Bin}(10, 0.254)$  and  $\Pr(Y \leq 2) = 0.514$ .

- c. Let  $\pi_{0.99}$  denote the 99-th percentile of this distribution. Then,

$$S_X(\pi_{0.99}) = \exp\left\{-\left(\frac{\pi_{0.99}}{33.33}\right)^{1.2}\right\} = 0.01.$$

Solving for  $\pi_{0.99}$ , we get  $\pi_{0.99} = 118.99$ .

#### 4.2.4 The Generalized Beta Distribution of the Second Kind

The Generalized Beta Distribution of the Second Kind (*GB2*) was introduced by [Venter \(1983\)](#) in the context of insurance loss modeling and by [McDonald \(1984\)](#) as an income and wealth distribution. It is a four-parameter, very flexible, distribution that can model positively as well as negatively skewed distributions.

The continuous variable  $X$  is said to have the *GB2* distribution with parameters  $\sigma$ ,  $\theta$ ,  $\alpha_1$  and  $\alpha_2$  if its *pdf* is given by

$$f_X(x) = \frac{(x/\theta)^{\alpha_2/\sigma}}{x\sigma B(\alpha_1, \alpha_2) \left[1 + (x/\theta)^{1/\sigma}\right]^{\alpha_1+\alpha_2}} \quad \text{for } x > 0, \quad (4.2)$$

$\sigma, \theta, \alpha_1, \alpha_2 > 0$ , and where the beta function  $B(\alpha_1, \alpha_2)$  is defined as

$$B(\alpha_1, \alpha_2) = \int_0^1 t^{\alpha_1-1} (1-t)^{\alpha_2-1} dt.$$

The *GB2* provides a model for heavy as well as light tailed data. It includes the exponential, gamma, Weibull, Burr, Lomax, F, chi-square, Rayleigh, log-normal and log-logistic as special or limiting cases. For example, by setting the parameters  $\sigma = \alpha_1 = \alpha_2 = 1$ , the *GB2* reduces to the log-logistic distribution. When  $\sigma = 1$  and  $\alpha_2 \rightarrow \infty$ , it reduces to the gamma distribution, and when  $\alpha = 1$  and  $\alpha_2 \rightarrow \infty$ , it reduces to the Weibull distribution.

A *GB2* random variable can be constructed as follows. Suppose that  $G_1$  and  $G_2$  are independent random variables where  $G_i$  has a gamma distribution with shape parameter  $\alpha_i$  and scale parameter 1. Then, one can show that the random variable  $X = \theta \left( \frac{G_1}{G_2} \right)^\sigma$  has a *GB2* distribution with *pdf* summarized in equation (4.2). This theoretical result has several implications. For example, when the moments exist, one can show that the  $k$ -th raw moment of the *GB2* distributed random variable is given by

$$E(X^k) = \frac{\theta^k B(\alpha_1 + k\sigma, \alpha_2 - k\sigma)}{B(\alpha_1, \alpha_2)}, \quad k > 0.$$

As will be described in Section 4.3.1, the *GB2* is also related to an *F*-distribution, a result that can be useful in simulation and residual analysis.

Earlier applications of the *GB2* were on income data and more recently have been used to model long-tailed claims data (Section ?? describes different interpretations of the descriptor “long-tail”). The *GB2* has been used to model different types of automobile insurance claims, severity of fire losses, as well as medical insurance claim data.

### 4.3 Methods of Creating New Distributions

In this section, you learn how to:

- Understand connections among the distributions
- Give insights into when a distribution is preferred when compared to alternatives
- Provide foundations for creating new distributions

#### 4.3.1 Functions of Random Variables and their Distributions

In Section 4.2 we discussed some elementary known distributions. In this section we discuss means of creating new parametric probability distributions from existing ones. Specifically, let  $X$  be a continuous random variable with a known *pdf*  $f_X(x)$  and distribution function  $F_X(x)$ . We are interested in the distribution of  $Y = g(X)$ , where  $g(X)$  is a one-to-one transformation defining a new random variable  $Y$ . In this section we apply the following techniques for creating new families of distributions: (a) multiplication by a constant (b) raising to a power, (c) exponentiation and (d) mixing.

##### Multiplication by a Constant

If claim data show change over time then such transformation can be useful to adjust for inflation. If the level of inflation is positive then claim costs are rising, and if it is negative then costs are falling. To adjust for inflation we multiply the cost  $X$  by 1+ inflation rate (negative inflation is deflation). To account for currency impact on claim costs we also use a transformation to apply currency conversion from a base to a counter currency.

Consider the transformation  $Y = cX$ , where  $c > 0$ , then the distribution function of  $Y$  is given by

$$F_Y(y) = \Pr(Y \leq y) = \Pr(cX \leq y) = \Pr\left(X \leq \frac{y}{c}\right) = F_X\left(\frac{y}{c}\right).$$

Using the chain rule for differentiation, the *pdf* of interest  $f_Y(y)$  can be written as

$$f_Y(y) = \frac{1}{c} f_X\left(\frac{y}{c}\right).$$

Suppose that  $X$  belongs to a certain set of parametric distributions and define a rescaled version  $Y = cX$ ,  $c > 0$ . If  $Y$  is in the same set of distributions

then the distribution is said to be a scale distribution. When a member of a scale distribution is multiplied by a constant  $c$  ( $c > 0$ ), the scale parameter for this scale distribution meets two conditions:

- The parameter is changed by multiplying by  $c$ ;
- All other parameters remain unchanged.

**Example 4.3.1. Actuarial Exam Question.** Losses of Eiffel Auto Insurance are denoted in Euro currency and follow a lognormal distribution with  $\mu = 8$  and  $\sigma = 2$ . Given that 1 euro = 1.3 dollars, find the set of lognormal parameters which describe the distribution of Eiffel's losses in dollars.

**Example Solution.** Let  $X$  and  $Y$  denote the aggregate losses of Eiffel Auto Insurance in euro currency and dollars respectively. As  $Y = 1.3X$ , we have,

$$F_Y(y) = \Pr(Y \leq y) = \Pr(1.3X \leq y) = \Pr\left(X \leq \frac{y}{1.3}\right) = F_X\left(\frac{y}{1.3}\right).$$

$X$  follows a lognormal distribution with parameters  $\mu = 8$  and  $\sigma = 2$ . The \*pdf\* of  $X$  is given by

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2\right\} \quad \text{for } x > 0.$$

As  $\left|\frac{dx}{dy}\right| = \frac{1}{1.3}$ , the \*pdf\* of interest  $f_Y(y)$  is

$$\begin{aligned} f_Y(y) &= \frac{1}{1.3} f_X\left(\frac{y}{1.3}\right) \\ &= \frac{1}{1.3} \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log(y/1.3)-\mu}{\sigma}\right)^2\right\} \\ &= \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log y - (\log 1.3 + \mu)}{\sigma}\right)^2\right\}. \end{aligned}$$

Then  $Y$  follows a lognormal distribution with parameters  $\log 1.3 + \mu = 8.26$  and  $\sigma = 2.00$ . If we let  $\mu = \log(m)$  then it can be easily seen that  $m = e^\mu$  is the scale parameter which was multiplied by 1.3 while  $\sigma$  is the shape parameter that remained unchanged.

---

**Example 4.3.2. Actuarial Exam Question.** Demonstrate that the gamma distribution is a scale distribution.

**Example Solution.** Let  $X \sim Ga(\alpha, \theta)$  and  $Y = cX$ . As  $\left| \frac{dx}{dy} \right| = \frac{1}{c}$ , then

$$f_Y(y) = \frac{1}{c} f_X\left(\frac{y}{c}\right) = \frac{\left(\frac{y}{c\theta}\right)^\alpha}{y \Gamma(\alpha)} \exp\left(-\frac{y}{c\theta}\right).$$

We can see that  $Y \sim Ga(\alpha, c\theta)$  indicating that gamma is a scale distribution and  $\theta$  is a scale parameter.

---

Using the same approach as in the example, you can demonstrate that other distributions introduced in Section 4.2 are also scale distributions. In actuarial modeling, working with a scale distribution is very convenient because it allows to incorporate the effect of inflation and to accommodate changes in the currency unit.

### Raising to a Power

In Section 4.2.3 we talked about the flexibility of the Weibull distribution in fitting reliability data. Looking to the origins of the Weibull distribution, we recognize that the Weibull is a power transformation of the exponential distribution. This is an application of another type of transformation which involves raising the random variable to a power.

Consider the transformation  $Y = X^\tau$ , where  $\tau > 0$ , then the distribution function of  $Y$  is given by

$$F_Y(y) = \Pr(Y \leq y) = \Pr(X^\tau \leq y) = \Pr(X \leq y^{1/\tau}) = F_X(y^{1/\tau}).$$

Hence, the *pdf* of interest  $f_Y(y)$  can be written as

$$f_Y(y) = \frac{1}{\tau} y^{(1/\tau)-1} f_X(y^{1/\tau}).$$

On the other hand, if  $\tau < 0$ , then the distribution function of  $Y$  is given by

$$F_Y(y) = \Pr(Y \leq y) = \Pr(X^\tau \leq y) = \Pr(X \geq y^{1/\tau}) = 1 - F_X(y^{1/\tau}),$$

and

$$f_Y(y) = \left| \frac{1}{\tau} \right| y^{(1/\tau)-1} f_X(y^{1/\tau}).$$

**Example 4.3.3.** We assume that  $X$  follows the exponential distribution with mean  $\theta$  and consider the transformed variable  $Y = X^\tau$ . Show that  $Y$  follows the Weibull distribution when  $\tau$  is positive and determine the parameters of the Weibull distribution.

**Example Solution.** As  $X$  follows the exponential distribution with mean  $\theta$ , we have

$$f_X(x) = \frac{1}{\theta} e^{-x/\theta} \quad x > 0.$$

Solving for  $x$  yields  $x = y^{1/\tau}$ . Taking the derivative, we have

$$\left| \frac{dx}{dy} \right| = \frac{1}{\tau} y^{\frac{1}{\tau}-1}.$$

Thus,

$$f_Y(y) = \frac{1}{\tau} y^{\frac{1}{\tau}-1} f_X\left(y^{\frac{1}{\tau}}\right) = \frac{1}{\tau\theta} y^{\frac{1}{\tau}-1} e^{-\frac{y^{\frac{1}{\tau}}}{\theta}} = \frac{\alpha}{\beta} \left(\frac{y}{\beta}\right)^{\alpha-1} e^{-(y/\beta)^\alpha}.$$

where  $\alpha = \frac{1}{\tau}$  and  $\beta = \theta^\tau$ . Then,  $Y$  follows the Weibull distribution with shape parameter  $\alpha$  and scale parameter  $\beta$ .

---

**Special Case. Relating a  $GB2$  to an  $F$ - Distribution.** We can use transforms such as multiplication by a constant and raising to a power to verify that the  $GB2$  distribution is related to an  $F$ -distribution, a distribution widely used in applied statistics.

---

To see this relationship, we first note that  $\frac{1}{2}G_1$  has a gamma distribution with shape parameter  $\alpha_1$  and scale parameter 0.5. Readers with some background in applied statistics may also recognize this to be a *chi-square* distribution with degrees of freedom  $2\alpha_1$ . The ratio of independent chi-squares has an  $F$ -distribution. That is

$$\frac{G_1}{G_2} = \frac{0.5G_1}{0.5G_2} = F$$

has an  $F$ -distribution with numerator degrees of freedom  $2\alpha_1$  and denominator degrees of freedom  $2\alpha_2$ . Thus, a random variable  $X$  with a  $GB2$  distribution can be expressed as  $X = \theta \left( \frac{G_1}{G_2} \right)^\sigma = \theta F^\sigma$ . With this, you can think of a  $GB2$  as a “power  $F$ ” or a “generalized  $F$ ”, as it is sometimes known in the literature.

Simulation, discussed in Chapter 8, provides a direct application of this result. Suppose we know how to simulate an outcome with an  $F$ -distribution (that is easy to do using, for example, the R function `rf(n,df1,df2)`), say  $F$ . Then we raise it to the power  $\sigma$  and multiply it by  $\theta$  so that  $\theta F^\sigma$  is an outcome that has a  $GB2$  distribution.

Residual analysis provides another direct application. Suppose we have an

outcome, say  $X$ , that we think comes from a  $GB2$  distribution. Then we can examine the transformed version  $X^* = (X/\theta)^{1/\sigma}$ . If the original specification is correct, then  $X^*$  has an  $F$ -distribution and there are many well-known techniques, some described in Chapter 6, for verifying this assertion.

---

### Exponentiation

The normal distribution is a very popular model for a wide number of applications and when the sample size is large, it can serve as an approximate distribution for other models. If the random variable  $X$  has a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , then  $Y = e^X$  has a lognormal distribution with parameters  $\mu$  and  $\sigma^2$ . The lognormal random variable has a lower bound of zero, is positively skewed and has a long right tail. A lognormal distribution is commonly used to describe distributions of financial assets such as stock prices. It is also used in fitting claim amounts for automobile as well as health insurance. This is an example of another type of transformation which involves exponentiation.

In general, consider the transformation  $Y = e^X$ . Then, the distribution function of  $Y$  is given by

$$F_Y(y) = \Pr(Y \leq y) = \Pr(e^X \leq y) = \Pr(X \leq \log y) = F_X(\log y).$$

Taking derivatives, we see that the *pdf* of interest  $f_Y(y)$  can be written as

$$f_Y(y) = \frac{1}{y} f_X(\log y).$$

As an important special case, suppose that  $X$  is normally distributed with mean  $\mu$  and variance  $\sigma^2$ . Then, the distribution of  $Y = e^X$  is

$$f_Y(y) = \frac{1}{y} f_X(\log y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log y - \mu}{\sigma}\right)^2\right\}.$$

This is known as a *lognormal* distribution.

**Example 4.3.4. Actuarial Exam Question.** Assume that  $X$  has a uniform distribution on the interval  $(0, c)$  and define  $Y = e^X$ . Find the distribution of  $Y$ .

**Example Solution.** We begin with the cdf of  $Y$ ,

$$F_Y(y) = \Pr(Y \leq y) = \Pr(e^X \leq y) = \Pr(X \leq \log y) = F_X(\log y).$$

Taking the derivative, we have,

$$f_Y(y) = \frac{1}{y} f_X(\log y) = \frac{1}{cy}.$$

Since  $0 < x < c$ , then  $1 < y < e^c$ .

### 4.3.2 Mixture Distributions for Severity

Mixture distributions represent a useful way of modeling data that are drawn from a heterogeneous population. This parent population can be thought to be divided into multiple subpopulations with distinct distributions.

#### Finite Mixtures

##### *Two-point Mixture*

If the underlying phenomenon is diverse and can actually be described as two phenomena representing two subpopulations with different modes, we can construct the two-point mixture random variable  $X$ . Given random variables  $X_1$  and  $X_2$ , with pdfs  $f_{X_1}(x)$  and  $f_{X_2}(x)$  respectively, the pdf of  $X$  is the weighted average of the component pdf  $f_{X_1}(x)$  and  $f_{X_2}(x)$ . The pdf and distribution function of  $X$  are given by

$$f_X(x) = af_{X_1}(x) + (1 - a)f_{X_2}(x),$$

and

$$F_X(x) = aF_{X_1}(x) + (1 - a)F_{X_2}(x),$$

for  $0 < a < 1$ , where the mixing parameters  $a$  and  $(1 - a)$  represent the proportions of data points that fall under each of the two subpopulations respectively. This weighted average can be applied to a number of other distribution related quantities. The  $k$ -th raw moment and moment generating function of  $X$  are given by  $E(X^k) = aE(X_1^k) + (1 - a)E(X_2^k)$ , and

$$M_X(t) = aM_{X_1}(t) + (1 - a)M_{X_2}(t),$$

respectively.

**Example 4.3.5. Actuarial Exam Question.** A collection of insurance policies consists of two types. 25% of policies are Type 1 and 75% of policies are Type 2. For a policy of Type 1, the loss amount per year follows an exponential distribution with mean 200, and for a policy of Type 2, the loss amount per year follows a Pareto distribution with parameters  $\alpha = 3$  and  $\theta = 200$ . For a policy chosen at random from the entire collection of both types of policies, find the probability that the annual loss will be less than 100, and find the average loss.

**Example Solution.** The two types of losses are the random variables  $X_1$  and  $X_2$ .  $X_1$  has an exponential distribution with mean 100, so  $F_{X_1}(100) = 1 - e^{-\frac{100}{200}} = 0.393$ .  $X_2$  has a Pareto distribution with parameters  $\alpha = 3$  and  $\theta = 200$ , so  $F_{X_2}(100) = 1 - \left(\frac{200}{100+200}\right)^3 = 0.704$ . Hence,  $F_X(100) = (0.25 \times 0.393) + (0.75 \times 0.704) = 0.626$ .

The average loss is given by

$$E(X) = 0.25E(X_1) + 0.75E(X_2) = (0.25 \times 200) + (0.75 \times 100) = 125.$$

#### k-point Mixture

In case of finite mixture distributions, the random variable of interest  $X$  has a probability  $p_i$  of being drawn from homogeneous subpopulation  $i$ , where  $i = 1, 2, \dots, k$  and  $k$  is the initially specified number of subpopulations in our mixture. The mixing parameter  $p_i$  represents the proportion of observations from subpopulation  $i$ . Consider the random variable  $X$  generated from  $k$  distinct subpopulations, where subpopulation  $i$  is modeled by the continuous distribution  $f_{X_i}(x)$ . The probability distribution of  $X$  is given by

$$f_X(x) = \sum_{i=1}^k p_i f_{X_i}(x),$$

where  $0 < p_i < 1$  and  $\sum_{i=1}^k p_i = 1$ .

This model is often referred to as a finite mixture or a  $k$ -point mixture. The distribution function,  $r$ -th raw moment and moment generating functions of the  $k$  point mixture are given as

$$F_X(x) = \sum_{i=1}^k p_i F_{X_i}(x),$$

$$E(X^r) = \sum_{i=1}^k p_i E(X_i^r), \quad \text{and}$$

$$M_X(t) = \sum_{i=1}^k p_i M_{X_i}(t),$$

respectively.

**Example 4.3.6. Actuarial Exam Question.**  $Y_1$  is a mixture of  $X_1$  and  $X_2$  with mixing weights  $a$  and  $(1 - a)$ .  $Y_2$  is a mixture of  $X_3$  and  $X_4$  with mixing

weights  $b$  and  $(1 - b)$ .  $Z$  is a mixture of  $Y_1$  and  $Y_2$  with mixing weights  $c$  and  $(1 - c)$ .

Show that  $Z$  is a mixture of  $X_1, X_2, X_3$  and  $X_4$ , and find the mixing weights.

**Example Solution.** Applying the formula for a mixed distribution, we get

$$f_{Y_1}(x) = af_{X_1}(x) + (1 - a)f_{X_2}(x)$$

$$f_{Y_2}(x) = bf_{X_3}(x) + (1 - b)f_{X_4}(x)$$

$$f_Z(x) = cf_{Y_1}(x) + (1 - c)f_{Y_2}(x).$$

Substituting the first two equations into the third, we get

$$f_Z(x) = c \left[ af_{X_1}(x) + (1 - a)f_{X_2}(x) \right] + (1 - c) \left[ bf_{X_3}(x) + (1 - b)f_{X_4}(x) \right]$$

$$= caf_{X_1}(x) + c(1 - a)f_{X_2}(x) + (1 - c)bf_{X_3}(x) + (1 - c)(1 - b)f_{X_4}(x).$$

Then,  $Z$  is a mixture of  $X_1, X_2, X_3$  and  $X_4$ , with mixing weights  $ca$ ,  $c(1 - a)$ ,  $(1 - c)b$  and  $(1 - c)(1 - b)$ , respectively. It can be easily seen that the mixing weights sum to one.

### Continuous Mixtures

A mixture with a very large number of subpopulations ( $k$  goes to infinity) is often referred to as a continuous mixture. In a continuous mixture, subpopulations are not distinguished by a discrete mixing parameter but by a continuous variable  $\Theta$ , where  $\Theta$  plays the role of  $p_i$  in the finite mixture. Consider the random variable  $X$  with a distribution depending on a parameter  $\Theta$ , where  $\Theta$  itself is a continuous random variable. This description yields the following model for  $X$

$$f_X(x) = \int_{-\infty}^{\infty} f_X(x|\theta) g_{\Theta}(\theta) d\theta,$$

where  $f_X(x|\theta)$  is the conditional distribution of  $X$  at a particular value of  $\Theta = \theta$  and  $g_{\Theta}(\theta)$  is the probability statement made about the unknown parameter  $\theta$ . In a Bayesian context (to be described in Chapter 9), this is known as the prior distribution of  $\Theta$  (the prior information or expert opinion to be used in the analysis).

The distribution function,  $k$ -th raw moment and moment generating functions of the continuous mixture are given as

$$F_X(x) = \int_{-\infty}^{\infty} F_X(x|\theta) g_{\Theta}(\theta) d\theta,$$

$$\begin{aligned} \mathbb{E}(X^k) &= \int_{-\infty}^{\infty} \mathbb{E}(X^k | \theta) g_{\Theta}(\theta) d\theta, \\ M_X(t) = \mathbb{E}(e^{tX}) &= \int_{-\infty}^{\infty} \mathbb{E}(e^{tx} | \theta) g_{\Theta}(\theta) d\theta, \end{aligned}$$

respectively.

The  $k$ -th raw moment of the mixture distribution can be rewritten as

$$\mathbb{E}(X^k) = \int_{-\infty}^{\infty} \mathbb{E}(X^k | \theta) g_{\Theta}(\theta) d\theta = \mathbb{E}[\mathbb{E}(X^k | \Theta)].$$

Using the law of iterated expectations (see Appendix Chapter ??), we can define the mean and variance of  $X$  as

$$\mathbb{E}(X) = \mathbb{E}[\mathbb{E}(X | \Theta)]$$

and

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | \Theta)] + \text{Var}[\mathbb{E}(X | \Theta)].$$

**Example 4.3.7. Actuarial Exam Question.**  $X$  has a normal distribution with a mean of  $\Lambda$  and variance of 1.  $\Lambda$  has a normal distribution with a mean of 1 and variance of 1. Find the mean and variance of  $X$ .

**Example Solution.**  $X$  is a continuous mixture with mean

$$\mathbb{E}(X) = \mathbb{E}[\mathbb{E}(X | \Lambda)] = \mathbb{E}(\Lambda) = 1$$

and

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X | \Lambda)] + \text{Var}[\mathbb{E}(X | \Lambda)] = \mathbb{E}(\text{Var}(X | \Lambda)) + \text{Var}(\mathbb{E}(X | \Lambda)) = \mathbb{E}(\text{Var}(X | \Lambda)) + \text{Var}(1) = \mathbb{E}(\text{Var}(X | \Lambda)) + 0 = \mathbb{E}(\text{Var}(X | \Lambda)).$$

**Example 4.3.8. Actuarial Exam Question.** Claim sizes,  $X$ , are uniform on the interval  $(\Theta, \Theta + 10)$  for each policyholder.  $\Theta$  varies by policyholder according to an exponential distribution with mean 5. Find the unconditional distribution, mean and variance of  $X$ .

**Example Solution.** The conditional distribution of  $X$  is  $f_X(x | \theta) = \frac{1}{10}$  for  $\theta < x < \theta + 10$ . The prior distribution of  $\theta$  is  $g_{\Theta}(\theta) = \frac{1}{5}e^{-\frac{\theta}{5}}$  for  $0 < \theta < \infty$ .

Multiplying and integrating yields the unconditional distribution of  $X$

$$f_X(x) = \int f_X(x | \theta) g_{\Theta}(\theta) d\theta.$$

For this example, this is

$$f_X(x) = \begin{cases} \int_0^x \frac{1}{50} e^{-\frac{\theta}{5}} d\theta = \frac{1}{10} (1 - e^{-\frac{x}{5}}) & 0 \leq x \leq 10, \\ \int_{x-10}^x \frac{1}{50} e^{-\frac{\theta}{5}} d\theta = \frac{1}{10} \left( e^{-\frac{(x-10)}{5}} - e^{-\frac{x}{5}} \right) & 10 < x < \infty. \end{cases}$$

One can use this to derive the mean and variance of the unconditional distribution. Alternatively, start with the conditional mean and variance of  $X$ , given by

$$E(X|\theta) = \frac{\theta + \theta + 10}{2} = \theta + 5$$

and

$$\text{Var}(X|\theta) = \frac{[(\theta + 10) - \theta]^2}{12} = \frac{100}{12},$$

respectively. With these, the unconditional mean and variance of  $X$  are given by

$$E(X) = E[E(X|\Theta)] = E(\Theta + 5) = E(\Theta) + 5 = 5 + 5 = 10,$$

and

$$\begin{aligned} \text{Var}(X) &= E[V(X|\Theta)] + \text{Var}[E(X|\Theta)] \\ &= E\left(\frac{100}{12}\right) + \text{Var}(\Theta + 5) = 8.33 + \text{Var}(\Theta) = 33.33. \end{aligned}$$

## 4.4 Estimating Loss Distributions

In this section, you learn how to:

- Estimate moments, quantiles, and distributions without reference to a parametric distribution
- Summarize the data graphically without reference to a parametric distribution
- Use method of moments, percentile matching, and maximum likelihood estimation to estimate parameters for different distributions.

### 4.4.1 Nonparametric Estimation

In Section 3.2 for frequency and Section 4.1 for severity, we learned how to summarize a distribution by computing means, variances, quantiles/percentiles, and so on. To approximate these summary measures using a dataset, one strategy is to:

- i. assume a parametric form for a distribution, such as a negative binomial for frequency or a gamma distribution for severity,
- ii. estimate the parameters of that distribution, and then
- iii. use the distribution with the estimated parameters to calculate the desired summary measure.

This is the parametric approach. Another strategy is to estimate the desired summary measure directly from the observations *without* reference to a parametric model. Not surprisingly, this is known as the nonparametric approach.

Let us start by considering the most basic type of sampling scheme and assume that observations are realizations from a set of random variables  $X_1, \dots, X_n$  that are iid draws from an unknown population distribution  $F(\cdot)$ . An equivalent way of saying this is that  $X_1, \dots, X_n$ , is a *random sample* (with replacement) from  $F(\cdot)$ . We now describe nonparametric estimators of many important measures that summarize a distribution.

### Moment Estimators

We learned how to define moments in Section 3.2.2 for frequency and Section 4.1.1 for severity. In particular, the  $k$ -th moment,  $E[X^k] = \mu'_k$ , summarizes many aspects of the distribution for different choices of  $k$ . Here,  $\mu'_k$  is sometimes called the  $k$ th *population* moment to distinguish it from the  $k$ th sample moment,

$$\frac{1}{n} \sum_{i=1}^n X_i^k,$$

which is the corresponding nonparametric estimator. In typical applications,  $k$  is a positive integer, although it need not be in theory. The sample estimator for the population mean  $\mu$  is called the *sample mean*, denoted with a bar on top of the random variable:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

A nonparametric, or sample, estimator of the  $k$ -th *central moment*,  $\mu_k$  is

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k.$$

Properties of the sample moment estimator of the variance such as  $n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$  have been studied extensively but is not the only possible estimator. The most widely used version is one where the effective sample

size is reduced by one, and so we define

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Dividing by  $n - 1$  instead of  $n$  matters little when you have a large sample size  $n$  as is common in insurance applications. The *sample variance* estimator  $s^2$  is unbiased in the sense that  $E[s^2] = \sigma^2$ , a desirable property particularly when interpreting results of an analysis.

### Empirical Distribution Function

We have seen how to compute nonparametric estimators of the  $k$ th moment  $E[X^k]$ . In the same way, for any known function  $g(\cdot)$ , we can estimate  $E[g(X)]$  using  $n^{-1} \sum_{i=1}^n g(X_i)$ .

Now consider the function  $g(X) = I(X \leq x)$  for a fixed  $x$ . Here, the notation  $I(\cdot)$  is the indicator function; it returns 1 if the event  $(\cdot)$  is true and 0 otherwise. Note that now the random variable  $g(X)$  has Bernoulli distribution (a binomial distribution with  $n = 1$ ). We can use this distribution to readily calculate quantities such as the mean and the variance. For example, for this choice of  $g(\cdot)$ , the expected value is  $E[I(X \leq x)] = \Pr(X \leq x) = F(x)$ , the distribution function evaluated at  $x$ . We define the nonparametric estimator of the distribution function

$$\begin{aligned} F_n(x) &= \frac{1}{n} \sum_{i=1}^n I(X_i \leq x) \\ &= \frac{\text{number of observations less than or equal to } x}{n}. \end{aligned}$$

As  $F_n(\cdot)$  is based on only observations and does not assume a parametric family for the distribution, it is nonparametric and also known as the empirical distribution function. It is also known as the *empirical cumulative distribution function* and, in R, one can use the `ecdf(.)` function to compute it.

**Example 4.4.1. Toy Data Set.** To illustrate, consider a fictitious, or “toy,” data set of  $n = 10$  observations. Determine the empirical distribution function.

$i$	1	2	3	4	5	6	7	8	9	10
$X_i$	10	15	15	15	20	23	23	23	23	30

You should check that the sample mean is  $\bar{X} = 19.7$  and that the sample variance is  $s^2 = 34.45556$ . The corresponding empirical distribution function

is

$$F_n(x) = \begin{cases} 0 & \text{for } x < 10 \\ 0.1 & \text{for } 10 \leq x < 15 \\ 0.4 & \text{for } 15 \leq x < 20 \\ 0.5 & \text{for } 20 \leq x < 23 \\ 0.9 & \text{for } 23 \leq x < 30 \\ 1 & \text{for } x \geq 30, \end{cases}$$

as shown in Figure 4.4. The empirical distribution is generally discrete and continuous from the right.

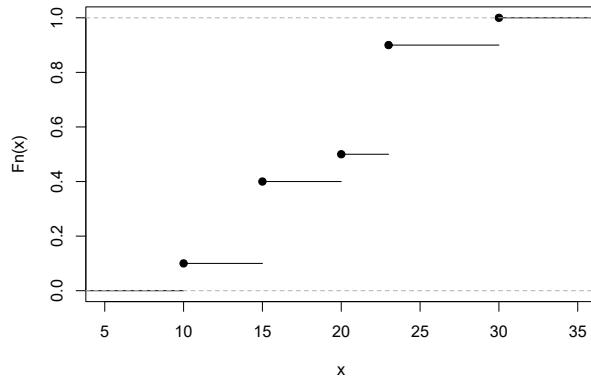


FIGURE 4.4: Empirical Distribution Function of a Toy Example

### Quartiles, Percentiles and Quantiles

We have already seen in Section 4.1.2 the median, which is the number such that approximately half of a data set is below (or above) it. The first quartile is the number such that approximately 25% of the data is below it and the third quartile is the number such that approximately 75% of the data is below it. A  $100p$  percentile is the number such that  $100 \times p$  percent of the data is below it.

To generalize this concept, consider a distribution function  $F(\cdot)$ , which may or may not be continuous, and let  $q$  be a fraction so that  $0 < q < 1$ . We want to define a quantile, say  $q_F$ , to be a number such that  $F(q_F) \approx q$ . Notice that when  $q = 0.5$ ,  $q_F$  is the median; when  $q = 0.25$ ,  $q_F$  is the first quartile, and so on. In the same way, when  $q = 0, 0.01, 0.02, \dots, 0.99, 1.00$ , the resulting  $q_F$  is a percentile. So, a quantile generalizes the concepts of median, quartiles, and percentiles.

To be precise, for a given  $0 < q < 1$ , define the  **$q$ th quantile  $q_F$**  to be *any*

number that satisfies

$$F(q_F-) \leq q \leq F(q_F) \quad (4.3)$$

Here, the notation  $F(x-)$  means to evaluate the function  $F(\cdot)$  as a left-hand limit.

To get a better understanding of this definition, let us look at a few special cases. First, consider the case where  $X$  is a continuous random variable so that the distribution function  $F(\cdot)$  has no jump points, as illustrated in Figure 4.5. In this figure, a few fractions,  $q_1$ ,  $q_2$ , and  $q_3$  are shown with their corresponding quantiles  $q_{F,1}$ ,  $q_{F,2}$ , and  $q_{F,3}$ . In each case, it can be seen that  $F(q_F-) = F(q_F)$  so that there is a unique quantile. Because we can find a unique inverse of the distribution function at any  $0 < q < 1$ , we can write  $q_F = F^{-1}(q)$ .

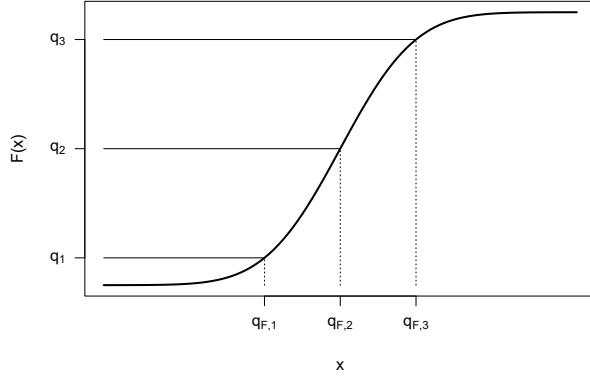


FIGURE 4.5: Continuous Quantile Case

Figure 4.6 shows three cases for distribution functions. The left panel corresponds to the continuous case just discussed. The middle panel displays a jump point similar to those we already saw in the empirical distribution function of Figure 4.4. For the value of  $q$  shown in this panel, we still have a unique value of the quantile  $q_F$ . Even though there are many values of  $q$  such that  $F(q_F-) \leq q \leq F(q_F)$ , for a particular value of  $q$ , there is only one solution to equation (4.3). The right panel depicts a situation in which the quantile cannot be uniquely determined for the  $q$  shown as there is a range of  $q_F$ 's satisfying equation (4.3).

**Example 4.4.2. Toy Data Set: Continued.** Determine quantiles corresponding to the 20th, 50th, and 95th percentiles.

**Solution.** Consider Figure 4.4. The case of  $q = 0.20$  corresponds to the middle

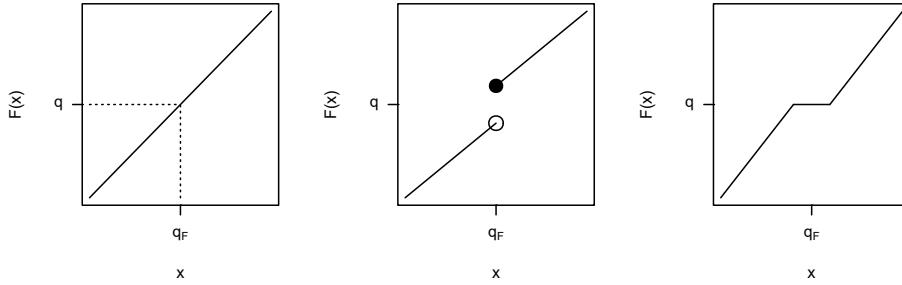


FIGURE 4.6: Three Quantile Cases

panel of Figure Figure 4.6, so the 20th percentile is 15. The case of  $q = 0.50$  corresponds to the right panel, so the median is any number between 20 and 23 inclusive. Many software packages use the average 21.5 (e.g. R, as seen below). For the 95th percentile, the solution is 30. We can see from Figure 4.4 that 30 also corresponds to the 99th and the 99.99th percentiles.

```
xExample <- c(10, rep(15, 3), 20, rep(23, 4), 30)
quantile(xExample, probs = c(0.2, 0.5, 0.95), type = 6)
```

---

```
20% 50% 95%
15.0 21.5 30.0
```

By taking a weighted average between data observations, smoothed empirical quantiles can handle cases such as the right panel in Figure 4.6. The  $q$ th smoothed empirical quantile is defined as

$$\hat{\pi}_q = (1 - h)X_{(j)} + hX_{(j+1)}$$

where  $j = \lfloor (n+1)q \rfloor$ ,  $h = (n+1)q - j$ , and  $X_{(1)}, \dots, X_{(n)}$  are the ordered values (known as the *order statistics*) corresponding to  $X_1, \dots, X_n$ . (Recall that the brackets  $\lfloor \cdot \rfloor$  are the floor function denoting the greatest integer value.) Note that  $\hat{\pi}_q$  is simply a linear interpolation between  $X_{(j)}$  and  $X_{(j+1)}$ .

**Example 4.4.3. Toy Data Set: Continued.** Determine the 50th and 20th smoothed percentiles.

**Example Solution.** Take  $n = 10$  and  $q = 0.5$ . Then,  $j = \lfloor (11)(0.5) \rfloor = \lfloor 5.5 \rfloor = 5$  and  $h = (11)(0.5) - 5 = 0.5$ . Then the 0.5-th smoothed empirical quantile is

$$\hat{\pi}_{0.5} = (1 - 0.5)X_{(5)} + (0.5)X_{(6)} = 0.5(20) + (0.5)(23) = 21.5.$$

Now take  $n = 10$  and  $q = 0.2$ . In this case,  $j = \lfloor (11)(0.2) \rfloor = \lfloor 2.2 \rfloor = 2$  and  $h = (11)(0.2) - 2 = 0.2$ . Then the 0.2-th smoothed empirical quantile is

$$\hat{\pi}_{0.2} = (1 - 0.2)X_{(2)} + (0.2)X_{(3)} = 0.8(15) + (0.2)(15) = 15.$$

### Density Estimators

**Discrete Variable.** When the random variable is discrete, estimating the probability mass function  $f(x) = \Pr(X = x)$  is straightforward. We simply use the sample average, defined to be

$$f_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i = x),$$

which is the proportion of the sample equal to  $x$ .

**Continuous Variable within a Group.** For a continuous random variable, consider a discretized formulation in which the domain of  $F(\cdot)$  is partitioned by constants  $\{c_0 < c_1 < \dots < c_k\}$  into intervals of the form  $[c_{j-1}, c_j]$ , for  $j = 1, \dots, k$ . The data observations are thus “grouped” by the intervals into which they fall. Then, we might use the basic definition of the empirical mass function, or a variation such as

$$f_n(x) = \frac{n_j}{n \times (c_j - c_{j-1})} \quad c_{j-1} \leq x < c_j,$$

where  $n_j$  is the number of observations ( $X_i$ ) that fall into the interval  $[c_{j-1}, c_j]$ .

**Continuous Variable (not grouped).** Extending this notion to instances where we observe individual data, note that we can always create arbitrary groupings and use this formula. More formally, let  $b > 0$  be a small positive constant, known as a bandwidth, and define a density estimator to be

$$f_n(x) = \frac{1}{2nb} \sum_{i=1}^n I(x - b < X_i \leq x + b) \quad (4.4)$$

---

**Snippet of Theory.** The idea is that the estimator  $f_n(x)$  in equation (4.4) is the average over  $n$  *iid* realizations of a random variable with mean

$$\begin{aligned} \mathbb{E} \left[ \frac{1}{2b} I(x - b < X \leq x + b) \right] &= \frac{1}{2b} (F(x + b) - F(x - b)) \\ &\rightarrow F'(x) = f(x), \end{aligned}$$

as  $b \rightarrow 0$ . That is,  $f_n(x)$  is an asymptotically unbiased estimator of  $f(x)$  (its

expectation approaches the true value as sample size increases to infinity). This development assumes some smoothness of  $F(\cdot)$ , in particular, twice differentiability at  $x$ , but makes no assumptions on the form of the distribution function  $F$ . Because of this, the density estimator  $f_n$  is said to be *nonparametric*.

More generally, define the kernel density estimator of the pdf at  $x$  as

$$f_n(x) = \frac{1}{nb} \sum_{i=1}^n w\left(\frac{x - X_i}{b}\right), \quad (4.5)$$

where  $w$  is a probability density function centered about 0. Note that equation (4.4) is a special case of the kernel density estimator where  $w(x) = \frac{1}{2}I(-1 < x \leq 1)$ , also known as the *uniform kernel*. Other popular choices are shown in Table 4.1.

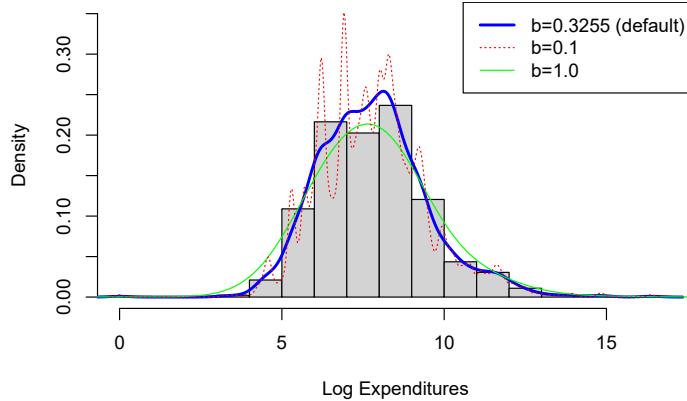
**Table 4.1. Popular Kernel Choices**

Kernel	$w(x)$
Uniform	$\frac{1}{2}I(-1 < x \leq 1)$
Triangle	$(1 -  x ) \times I( x  \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - x^2) \times I( x  \leq 1)$
Gaussian	$\phi(x)$

Here,  $\phi(\cdot)$  is the standard normal density function. As we will see in the following example, the choice of bandwidth  $b$  comes with a bias-variance tradeoff between matching local distributional features and reducing the volatility.

**Example 4.4.4. Property Fund.** Figure 4.7 shows a histogram (with shaded gray rectangles) of logarithmic property claims from 2010. The (blue) thick curve represents a Gaussian kernel density where the bandwidth was selected automatically using an ad hoc rule based on the sample size and volatility of these data. For this dataset, the bandwidth turned out to be  $b = 0.3255$ . For comparison, the (red) dashed curve represents the density estimator with a bandwidth equal to 0.1 and the green smooth curve uses a bandwidth of 1. As anticipated, the smaller bandwidth (0.1) indicates taking local averages over less data so that we get a better idea of the local average, but at the price of higher volatility. In contrast, the larger bandwidth (1) smooths out local fluctuations, yielding a smoother curve that may miss perturbations in the local average. For actuarial applications, we mainly use the kernel density estimator to get a quick visual impression of the data. From this perspective,

you can simply use the default ad hoc rule for bandwidth selection, knowing that you have the ability to change it depending on the situation at hand.



**FIGURE 4.7: Histogram of Logarithmic Property Claims with Superimposed Kernel Density Estimators**

---

Nonparametric density estimators, such as the kernel estimator, are regularly used in practice. The concept can also be extended to give smooth versions of an empirical distribution function. Given the definition of the kernel density estimator, the *kernel estimator of the distribution function* can be found as

$$\tilde{F}_n(x) = \frac{1}{n} \sum_{i=1}^n W\left(\frac{x - X_i}{b}\right).$$

where  $W$  is the distribution function associated with the kernel density  $w$ . To illustrate, for the uniform kernel, we have  $w(y) = \frac{1}{2}I(-1 < y \leq 1)$ , so

$$W(y) = \begin{cases} 0 & y < -1 \\ \frac{y+1}{2} & -1 \leq y < 1 \\ 1 & y \geq 1 \end{cases}$$

---

**Example 4.4.5. Actuarial Exam Question.** You study five lives to estimate the time from the onset of a disease to death. The times to death are:

2   3   3   3   7

Using a triangular kernel with bandwidth 2, calculate the density function estimate at 2.5.

**Example Solution.** For the kernel density estimate, we have

$$f_n(x) = \frac{1}{nb} \sum_{i=1}^n w\left(\frac{x - X_i}{b}\right),$$

where  $n = 5$ ,  $b = 2$ , and  $x = 2.5$ . For the triangular kernel,  $w(x) = (1 - |x|) \times I(|x| \leq 1)$ . Thus,

$X_i$	$\frac{x - X_i}{b}$	$w\left(\frac{x - X_i}{b}\right)$
2	$\frac{2.5-2}{2} = \frac{1}{4}$	$(1 - \frac{1}{4})(1) = \frac{3}{4}$
3		
3	$\frac{2.5-3}{2} = -\frac{1}{4}$	$(1 -  - \frac{1}{4} )(1) = \frac{3}{4}$
3		
7	$\frac{2.5-7}{2} = -2.25$	$(1 -  -2.25 )(0) = 0$

Then the kernel density estimate at  $x = 2.5$  is

$$f_n(2.5) = \frac{1}{5(2)} \left( \frac{3}{4} + (3)\frac{3}{4} + 0 \right) = \frac{3}{10}.$$

#### 4.4.2 Parametric Estimation

Section 4.2 has focused on parametric distributions that are commonly used in insurance applications. However, to be useful in applied work, these distributions must use “realistic” values for the parameters. In this section we cover three methods for estimating parameters: Method of moments, Percentile matching, and Maximum likelihood estimation.

##### Method of Moments

Under the method of moments, we approximate the moments of the parametric distribution using the empirical moments described in Section 4.4.1. We can then algebraically solve for the parameter estimates.

---

**Example 4.4.6. Property Fund.** For the 2010 property fund, there are  $n = 1,377$  individual claims (in thousands of dollars) with

$$m_1 = \frac{1}{n} \sum_{i=1}^n X_i = 26.62259 \quad \text{and} \quad m_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 = 136154.6.$$

Fit the parameters of the gamma and Pareto distributions using the method of moments.

**Solution.** To fit a gamma distribution, we have  $\mu_1 = \alpha\theta$  and  $\mu'_2 = \alpha(\alpha+1)\theta^2$ . Equating the two yields the method of moments estimators, easy algebra shows that

$$\alpha = \frac{\mu_1^2}{\mu'_2 - \mu_1^2} \quad \text{and} \quad \theta = \frac{\mu'_2 - \mu_1^2}{\mu_1}.$$

Thus, the method of moment estimators are

$$\hat{\alpha} = \frac{26.62259^2}{136154.6 - 26.62259^2} = 0.005232809$$

$$\hat{\theta} = \frac{136154.6 - 26.62259^2}{26.62259} = 5,087.629.$$

For comparison, the maximum likelihood values (see Section 4.4.2) turn out to be  $\hat{\alpha}_{MLE} = 0.2905959$  and  $\hat{\theta}_{MLE} = 91.61378$ , so there are big discrepancies between the two estimation procedures. This is one indication, as we have seen before, that the gamma model fits poorly.

In contrast, now assume a Pareto distribution so that  $\mu_1 = \theta/(\alpha - 1)$  and  $\mu'_2 = 2\theta^2/((\alpha - 1)(\alpha - 2))$ . Note that this expression for  $\mu'_2$  is only valid for  $\alpha > 2$ . Easy algebra shows

$$\alpha = 1 + \frac{\mu'_2}{\mu'_2 - \mu_1^2} \quad \text{and} \quad \theta = (\alpha - 1)\mu_1.$$

Thus, the method of moment estimators are

$$\hat{\alpha} = 1 + \frac{136154.6}{136154.6 - 26,62259^2} = 2.005233$$

$$\hat{\theta} = (2.005233 - 1) \cdot 26.62259 = 26.7619.$$

The maximum likelihood values turn out to be  $\hat{\alpha}_{MLE} = 0.9990936$  and  $\hat{\theta}_{MLE} = 2.2821147$ . It is interesting that  $\hat{\alpha}_{MLE} < 1$ ; for the Pareto distribution, recall that  $\alpha < 1$  means that the mean is infinite. This is another indication that the property claims data set is a long tail distribution.

As the above example suggests, there is flexibility with the method of moments. For example, we could have matched the second and third moments instead of the first and second, yielding different estimators. Furthermore, there is no guarantee that a solution will exist for each problem. For data that are censored or truncated, matching moments is possible for a few problems but, in general, this is a more difficult scenario. Finally, for distributions where the moments do not exist or are infinite, method of moments is not available. As an alternative, one can use the percentile matching technique.

### Percentile Matching

Under percentile matching, we approximate the quantiles or percentiles of the parametric distribution using the empirical quantiles or percentiles described in Section 4.4.1.

---

**Example 4.4.7. Property Fund.** For the 2010 property fund, we illustrate matching on quantiles. In particular, the Pareto distribution is intuitively pleasing because of the closed-form solution for the quantiles. Recall that the distribution function for the Pareto distribution is

$$F(x) = 1 - \left( \frac{\theta}{x + \theta} \right)^{\alpha}.$$

Easy algebra shows that we can express the quantile as

$$F^{-1}(q) = \theta \left( (1 - q)^{-1/\alpha} - 1 \right),$$

for a fraction  $q$ ,  $0 < q < 1$ .

Determine estimates of the Pareto distribution parameters using the 25th and 95th empirical quantiles.

**Example Solution.** The 25th percentile (the first quartile) turns out to be 0.78853 and the 95th percentile is 50.98293 (both in thousands of dollars). With two equations

$$0.78853 = \theta \left( 1 - (1 - .25)^{-1/\alpha} \right) \quad \text{and} \quad 50.98293 = \theta \left( 1 - (1 - .75)^{-1/\alpha} \right)$$

and two unknowns, the solution is  $\hat{\alpha} = 0.9412076$  and  $\hat{\theta} = 2.205617$ .

We remark here that a numerical routine is required for these solutions as no analytic solution is available. Furthermore, recall that the maximum likelihood estimates are  $\hat{\alpha}_{MLE} = 0.9990936$  and  $\hat{\theta}_{MLE} = 2.2821147$ , so the percentile matching provides a better approximation for the Pareto distribution than the method of moments.

---

**Example 4.4.8. Actuarial Exam Question.** You are given:

- (i) Losses follow a loglogistic distribution with cumulative distribution function:

$$F(x) = \frac{(x/\theta)^{\gamma}}{1 + (x/\theta)^{\gamma}}$$

(ii) The sample of losses is:

10 35 80 86 90 120 158 180 200 210 1500

Calculate the estimate of  $\theta$  by percentile matching, using the 40th and 80th empirically smoothed percentile estimates.

**Example Solution.** With 11 observations, we have  $j = \lfloor (n+1)q \rfloor = \lfloor 12(0.4) \rfloor = \lfloor 4.8 \rfloor = 4$  and  $h = (n+1)q - j = 12(0.4) - 4 = 0.8$ . By interpolation, the 40th empirically smoothed percentile estimate is  $\hat{\pi}_{0.4} = (1-h)X_{(j)} + hX_{(j+1)} = 0.2(86) + 0.8(90) = 89.2$ .

Similarly, for the 80th empirically smoothed percentile estimate, we have  $12(0.8) = 9.6$  so the estimate is  $\hat{\pi}_{0.8} = 0.4(200) + 0.6(210) = 206$ .

Using the loglogistic cumulative distribution, we need to solve the following two equations for parameters  $\hat{\theta}$  and  $\hat{\gamma}$ :

$$0.4 = \frac{(89.2/\hat{\theta})^{\hat{\gamma}}}{1 + (89.2/\hat{\theta})^{\hat{\gamma}}} \quad \text{and} \quad 0.8 = \frac{(206/\hat{\theta})^{\hat{\gamma}}}{1 + (206/\hat{\theta})^{\hat{\gamma}}}.$$

Solving for each parenthetical expression gives  $\frac{2}{3} = (89.2/\hat{\theta})^{\hat{\gamma}}$  and  $4 = (206/\hat{\theta})^{\hat{\gamma}}$ . Taking the ratio of the second equation to the first gives  $6 = (206/89.2)^{\hat{\gamma}} \Rightarrow \hat{\gamma} = \frac{\log(6)}{\log(206/89.2)} = 2.1407$ . Then  $4^{1/2.1407} = 206/\hat{\theta} \Rightarrow \hat{\theta} = 107.8$ .

---

Like the method of moments, percentile matching is almost too flexible in the sense that estimators can vary depending on different percentiles chosen. For example, one actuary may use estimation on the 25th and 95th percentiles whereas another uses the 20th and 80th percentiles. In general estimated parameters will differ and there is no compelling reason to prefer one over the other. Also as with the method of moments, percentile matching is appealing because it provides a technique that can be readily applied in selected situations and has an intuitive basis. Although most actuarial applications use maximum likelihood estimators, it can be convenient to have alternative approaches such as method of moments and percentile matching available.

---

#### Maximum Likelihood Estimators for Complete Data

At a foundational level, we assume that the analyst has available a random sample  $X_1, \dots, X_n$  from a distribution with distribution function  $F_X$  (for brevity, we sometimes drop the subscript  $X$ ). As is common, we use the vector  $\theta$  to denote the set of parameters for  $F$ . This basic sample scheme is reviewed

in Appendix Section ???. Although basic, this sampling scheme provides the foundations for understanding more complex schemes that are regularly used in practice, and so it is important to master the basics.

Before drawing from a distribution, we consider potential outcomes summarized by the random variable  $X_i$  (here,  $i$  is 1, 2, ...,  $n$ ). After the draw, we observe  $x_i$ . Notationally, we use uppercase roman letters for random variables and lower case ones for realizations. We have seen this set-up already in Section 3.4, where we used  $\Pr(X_1 = x_1, \dots, X_n = x_n)$  to quantify the “likelihood” of drawing a sample  $\{x_1, \dots, x_n\}$ . With continuous data, we use the joint probability density function instead of joint probabilities. With the independence assumption, the joint *pdf* may be written as the product of pdfs. Thus, we define the **likelihood** to be

$$L(\theta) = \prod_{i=1}^n f(x_i). \quad (4.6)$$

From the notation, note that we consider this to be a function of the parameters in  $\theta$ , with the data  $\{x_1, \dots, x_n\}$  held fixed. The maximum likelihood estimator is that value of the parameters in  $\theta$  that maximize  $L(\theta)$ .

From calculus, we know that maximizing a function produces the same results as maximizing the logarithm of a function (this is because the logarithm is a monotone function). Because we get the same results, to ease computational considerations, it is common to consider the **logarithmic likelihood**, denoted as

$$l(\theta) = \log L(\theta) = \sum_{i=1}^n \log f(x_i). \quad (4.7)$$

Appendix Section ?? reviews the foundations of maximum likelihood estimation with more mathematical details in Appendix Chapter ??.

**Example 4.4.9. Actuarial Exam Question.** You are given the following five observations: 521, 658, 702, 819, 1217. You use the single-parameter Pareto with distribution function:

$$F(x) = 1 - \left( \frac{500}{x} \right)^\alpha, \quad x > 500.$$

With  $n = 5$ , the log-likelihood function is

$$l(\alpha) = \sum_{i=1}^5 \log f(x_i; \alpha) = 5\alpha \log 500 + 5 \log \alpha - (\alpha + 1) \sum_{i=1}^5 \log x_i.$$

Figure 4.8 shows the logarithmic likelihood as a function of the parameter  $\alpha$ .

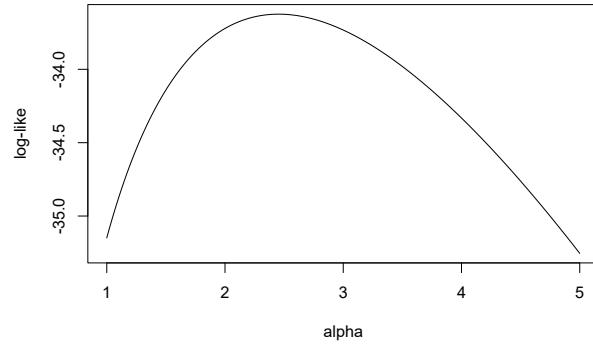


FIGURE 4.8: Logarithmic Likelihood for a One-Parameter Pareto

We can determine the maximum value of the logarithmic likelihood by taking derivatives and setting it equal to zero. This yields

$$\begin{aligned}\frac{\partial}{\partial \alpha} l(\alpha) &= 5 \log 500 + 5/\alpha - \sum_{i=1}^5 \log x_i =_{set} 0 \Rightarrow \\ \hat{\alpha}_{MLE} &= \frac{5}{\sum_{i=1}^5 \log x_i - 5 \log 500} = 2.453.\end{aligned}$$

Naturally, there are many problems where it is not practical to use hand calculations for optimization. Fortunately there are many statistical routines available such as the R function `optim`.

This code confirms our hand calculation result where the maximum likelihood estimator is  $\hat{\alpha}_{MLE} = 2.453125$ .

We present a few additional examples to illustrate how actuaries fit a parametric distribution model to a set of claim data using maximum likelihood.

**Example 4.4.10. Actuarial Exam Question.** Consider a random sample of claim amounts: 8000 10000 12000 15000. You assume that claim amounts follow an inverse exponential distribution, with parameter  $\theta$ . Calculate the maximum likelihood estimator for  $\theta$ .

**Example Solution.** The *pdf* is

$$f_X(x) = \frac{\theta e^{-\frac{\theta}{x}}}{x^2},$$

where  $x > 0$ .

The likelihood function,  $L(\theta)$ , can be viewed as the probability of the observed data, written as a function of the model's parameter  $\theta$

$$L(\theta) = \prod_{i=1}^4 f_{X_i}(x_i) = \frac{\theta^4 e^{-\theta \sum_{i=1}^4 \frac{1}{x_i}}}{\prod_{i=1}^4 x_i^2}.$$

The log-likelihood function,  $\log L(\theta)$ , is the sum of the individual logarithms

$$\log L(\theta) = 4 \log \theta - \theta \sum_{i=1}^4 \frac{1}{x_i} - 2 \sum_{i=1}^4 \log x_i.$$

Taking a derivative, we have

$$\frac{d \log L(\theta)}{d\theta} = \frac{4}{\theta} - \sum_{i=1}^4 \frac{1}{x_i}.$$

The maximum likelihood estimator of  $\theta$ , denoted by  $\hat{\theta}$ , is the solution to the equation

$$\frac{4}{\hat{\theta}} - \sum_{i=1}^4 \frac{1}{x_i} = 0.$$

Thus,  $\hat{\theta} = \frac{4}{\sum_{i=1}^4 \frac{1}{x_i}} = 10,667$ .

The second derivative of  $\log L(\theta)$  is given by

$$\frac{d^2 \log L(\theta)}{d\theta^2} = \frac{-4}{\theta^2}.$$

Evaluating the second derivative of the loglikelihood function at  $\hat{\theta} = 10,667$  gives a negative value, indicating  $\hat{\theta}$  as the value that maximizes the loglikelihood function.

---

**Example 4.4.11. Actuarial Exam Question.** A random sample of size 6 is from a lognormal distribution with parameters  $\mu$  and  $\sigma$ . The sample values are

200    3000    8000    60000    60000    160000.

Calculate the maximum likelihood estimator for  $\mu$  and  $\sigma$ .

**Example Solution.** The *pdf* is

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2\right),$$

where  $x > 0$ .

The likelihood function,  $L(\mu, \sigma)$ , is the product of the *pdf* for each data point.

$$L(\mu, \sigma) = \prod_{i=1}^6 f_{X_i}(x_i) = \frac{1}{\sigma^6 (2\pi)^3 \prod_{i=1}^6 x_i} \exp\left(-\frac{1}{2} \sum_{i=1}^6 \left(\frac{\log x_i - \mu}{\sigma}\right)^2\right).$$

Taking a logarithm yields the loglikelihood function,  $\log L(\mu, \sigma)$ , which is the sum of the individual logarithms.

$$\log L(\mu, \sigma) = -6 \log \sigma - 3 \log(2\pi) - \sum_{i=1}^6 \log x_i - \frac{1}{2} \sum_{i=1}^6 \left(\frac{\log x_i - \mu}{\sigma}\right)^2.$$

The first partial derivatives are

$$\begin{aligned} \frac{\partial \log L(\mu, \sigma)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^6 (\log x_i - \mu) \\ \frac{\partial \log L(\mu, \sigma)}{\partial \sigma} &= \frac{-6}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^6 (\log x_i - \mu)^2. \end{aligned}$$

The maximum likelihood estimators of  $\mu$  and  $\sigma$ , denoted by  $\hat{\mu}$  and  $\hat{\sigma}$ , are the solutions to the equations

$$\begin{aligned} \frac{1}{\hat{\sigma}^2} \sum_{i=1}^6 (\log x_i - \hat{\mu}) &= 0 \\ \frac{-6}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3} \sum_{i=1}^6 (\log x_i - \hat{\mu})^2 &= 0. \end{aligned}$$

These yield the estimates

$$\hat{\mu} = \frac{\sum_{i=1}^6 \log x_i}{6} = 9.38 \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^6 (\log x_i - \hat{\mu})^2}{6} = 5.12.$$

To check that these estimates maximize, and do not minimize, the likelihood, you may also wish to compute the second partial derivatives. These are

$$\frac{\partial^2 \log L(\mu, \sigma)}{\partial \mu^2} = \frac{-6}{\sigma^2}, \quad \frac{\partial^2 \log L(\mu, \sigma)}{\partial \mu \partial \sigma} = \frac{-2}{\sigma^3} \sum_{i=1}^6 (\log x_i - \mu)$$

and

$$\frac{\partial^2 \log L(\mu, \sigma)}{\partial \sigma^2} = \frac{6}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^6 (\log x_i - \mu)^2.$$

Two follow-up questions rely on large sample properties that you may have seen in an earlier course. Appendix Chapter ?? reviews the definition of the likelihood function, introduces its properties, reviews the maximum likelihood estimators, extends their large-sample properties to the case where there are multiple parameters in the model, and reviews statistical inference based on

maximum likelihood estimators. In the solutions of these examples we derive the asymptotic variance of maximum-likelihood estimators of the model parameters. We use the delta method to derive the asymptotic variances of functions of these parameters.

**Example 4.4.10 - Follow - Up.** Refer to **Example 4.4.10.**

- Approximate the variance of the maximum likelihood estimator.
- Determine an approximate 95% confidence interval for  $\theta$ .
- Determine an approximate 95% confidence interval for  $\Pr(X \leq 9,000)$ .

**Example Solution.**

a. Taking reciprocal of negative expectation of the second derivative of  $\log L(\theta)$ , we obtain an estimate of the variance of  $\hat{\theta}$ ,  $\widehat{Var}(\hat{\theta}) = \left[ E\left(\frac{d^2 \log L(\theta)}{d\theta^2}\right) \right]^{-1} \Big|_{\theta=\hat{\theta}} = \frac{\hat{\theta}^2}{4} = 28,446,222$ .

It should be noted that as the sample size  $n \rightarrow \infty$ , the distribution of the maximum likelihood estimator  $\hat{\theta}$  converges to a normal distribution with mean  $\theta$  and variance  $\widehat{V}(\hat{\theta})$ . The approximate confidence interval in this example is based on the assumption of normality, despite the small sample size, only for the purpose of illustration.

b. The 95

$$10,667 \pm 1.96\sqrt{28,446,222} = (213.34, 21120.66).$$

c. The distribution function of  $X$  is  $F(x) = 1 - e^{-\frac{x}{\theta}}$ . Then, the maximum likelihood estimate of  $g_\Theta(\theta) = F(9,000)$  is

$$g(\hat{\theta}) = 1 - e^{-\frac{9,000}{10,667}} = 0.57.$$

We use the delta method to approximate the variance of  $g(\hat{\theta})$ .

$$\frac{dg(\theta)}{d\theta} = -\frac{9000}{\theta^2} e^{-\frac{9000}{\theta}}.$$

$$\widehat{Var}[g(\hat{\theta})] = \left(-\frac{9000}{\hat{\theta}^2} e^{-\frac{9000}{\hat{\theta}}}\right)^2 \widehat{V}(\hat{\theta}) = 0.0329.$$

The 95

$$0.57 \pm 1.96\sqrt{0.0329} = (0.214, 0.926).$$

**Example 4.4.11 - Follow - Up.** Refer to **Example 4.4.11.**

- Estimate the covariance matrix of the maximum likelihood estimator.
- Determine approximate 95% confidence intervals for  $\mu$  and  $\sigma$ .
- Determine an approximate 95% confidence interval for the mean of the lognormal distribution.

**a.** To derive the covariance matrix of the mle we need to find the expectations of the second derivatives. Since the random variable  $X$  is from a lognormal distribution with parameters  $\mu$  and  $\sigma$ , then  $\log X$  is normally distributed with mean  $\mu$  and variance  $\sigma^2$ .

$$E\left(\frac{\partial^2 \log L(\mu, \sigma)}{\partial \mu^2}\right) = E\left(\frac{-6}{\sigma^2}\right) = \frac{-6}{\sigma^2},$$

$$\begin{aligned} E\left(\frac{\partial^2 \log L(\mu, \sigma)}{\partial \mu \partial \sigma}\right) &= \frac{-2}{\sigma^3} \sum_{i=1}^6 E(\log x_i - \mu) \\ &= \frac{-2}{\sigma^3} \sum_{i=1}^6 [E(\log x_i) - \mu] = \frac{-2}{\sigma^3} \sum_{i=1}^6 (\mu - \mu) = 0, \end{aligned}$$

and

$$\begin{aligned} E\left(\frac{\partial^2 \log L(\mu, \sigma)}{\partial \sigma^2}\right) &= \frac{6}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^6 E(\log x_i - \mu)^2 \\ &= \frac{6}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^6 \text{Var}(\log x_i) = \frac{6}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^6 \sigma^2 \\ &= \frac{12}{\sigma^2}. \end{aligned}$$

Using the negatives of these expectations we obtain the Fisher information matrix

$$\begin{bmatrix} \frac{6}{\sigma^2} & 0 \\ 0 & \frac{12}{\sigma^2} \end{bmatrix}.$$

The covariance matrix,  $\Sigma$ , is the inverse of the Fisher information matrix

$$\Sigma = \begin{bmatrix} \frac{\sigma^2}{6} & 0 \\ 0 & \frac{\sigma^2}{12} \end{bmatrix}.$$

The estimated matrix is given by

$$\hat{\Sigma} = \begin{bmatrix} 0.8533 & 0 \\ 0 & 0.4267 \end{bmatrix}.$$

**b.** The 95% confidence interval for  $\mu$  is given by  $9.38 \pm 1.96\sqrt{0.8533} = (7.57, 11.19)$ .

The 95% confidence interval for  $\sigma^2$  is given by  $5.12 \pm 1.96\sqrt{0.4267} = (3.84, 6.40)$ .

c. The mean of  $X$  is  $\exp\left(\mu + \frac{\sigma^2}{2}\right)$ . Then, the maximum likelihood estimate of

$$g(\mu, \sigma) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

is

$$g(\hat{\mu}, \hat{\sigma}) = \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right) = 153,277.$$

We use the delta method to approximate the variance of the mle  $g(\hat{\mu}, \hat{\sigma})$ .

$$\frac{\partial g(\mu, \sigma)}{\partial \mu} = \exp\left(\mu + \frac{\sigma^2}{2}\right) \text{ and } \frac{\partial g(\mu, \sigma)}{\partial \sigma} = \sigma \exp\left(\mu + \frac{\sigma^2}{2}\right).$$

Using the delta method, the approximate variance of  $g(\hat{\mu}, \hat{\sigma})$  is given by

$$\begin{aligned}\widehat{\text{Var}}(g(\hat{\mu}, \hat{\sigma})) &= \begin{bmatrix} \frac{\partial g(\mu, \sigma)}{\partial \mu} & \frac{\partial g(\mu, \sigma)}{\partial \sigma} \end{bmatrix} \Sigma \begin{bmatrix} \frac{\partial g(\mu, \sigma)}{\partial \mu} \\ \frac{\partial g(\mu, \sigma)}{\partial \sigma} \end{bmatrix} \Big|_{\mu=\hat{\mu}, \sigma=\hat{\sigma}} \\ &= [153,277 \quad 346,826] \begin{bmatrix} 0.8533 & 0 \\ 0 & 0.4267 \end{bmatrix} \begin{bmatrix} 153,277 \\ 346,826 \end{bmatrix} \\ &= 71,374,380,000.\end{aligned}$$

The 95% confidence interval for  $\exp\left(\mu + \frac{\sigma^2}{2}\right)$  is given by

$$153277 \pm 1.96\sqrt{71,374,380,000} = (-370356, 676910).$$

Since the mean of the lognormal distribution cannot be negative, we should replace the negative lower limit in the previous interval by a zero.

**Example 4.4.12. Wisconsin Property Fund.** To see how maximum likelihood estimators work with real data, we return to the 2010 claims data introduced in Section 1.3.

The following snippet of code shows how to fit the exponential, gamma, Pareto, lognormal, and *GB2* models. For consistency, the code employs the R package **VGAM**. The acronym stands for *Vector Generalized Linear and Additive Models*; as suggested by the name, this package can do far more than fit these models although it suffices for our purposes. The one exception is the *GB2* density which is not widely used outside of insurance applications; however, we can code this density and compute maximum likelihood estimators using the **optim** general purpose optimizer.

Results from the fitting exercise are summarized in Figure 4.9. Here, the black

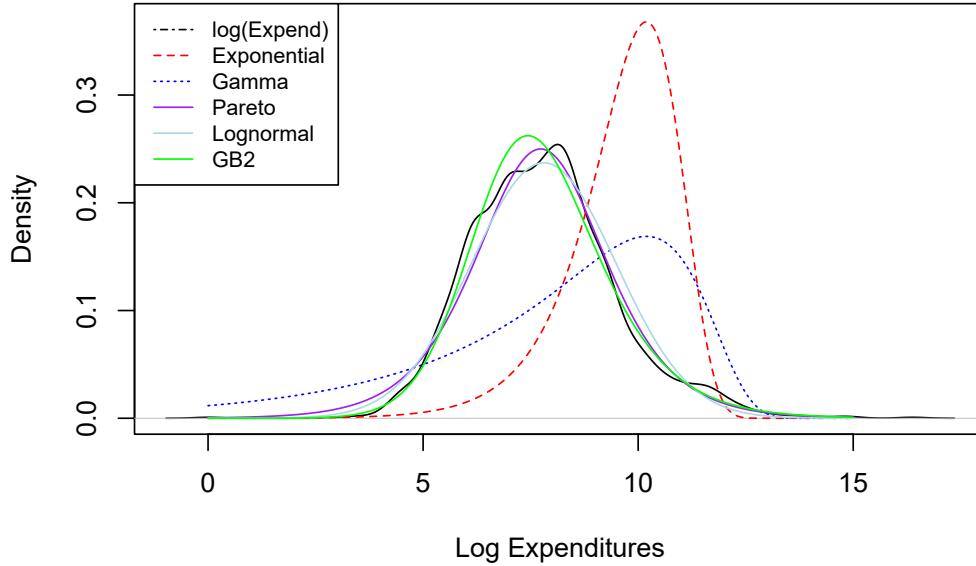


FIGURE 4.9: Density Comparisons for the Wisconsin Property Fund

“longdash” curve is a density estimator of the actual data (introduced in Section 4.4.1); the other curves are parametric curves where the parameters are computed via maximum likelihood. We see poor fits in the red dashed line from the exponential distribution fit and the blue dotted line from the gamma distribution fit. Fits of the other curves, Pareto, lognormal, and GB2, all seem to provide reasonably good fits to the actual data. Chapter 6 describes in more detail the principles of model selection.

### Starting Values

Generally, maximum likelihood is the preferred technique for parameter estimation because it employs data more efficiently. (See Appendix Chapter ?? for precise definitions of efficiency.) However, methods of moments and percentile matching are useful because they are easier to interpret and therefore allow the actuary or analyst to explain procedures to others. Additionally, the numerical estimation procedure (e.g. if performed in R) for the maximum likelihood is iterative and requires starting values to begin the recursive process. Although many problems are robust to the choice of the starting values, for some complex situations it can be important to have a starting value that

is close to the (unknown) optimal value. Method of moments and percentile matching can produce desirable estimates without a serious computational investment and can thus be used as a *starting value* for computing maximum likelihood.

---

## 4.5 Exercises with a Practical Focus

**Exercise 4.1. Corporate Travel** This exercise is based on the data set introduced in [Exercise 1.1](#) where now the focus is on severity modeling. As in [Exercise 3.14](#), we fit data for the period 2006-2021 but restrict claims to be greater than or equal to 10 (Australian dollars).

- a. Using the R function `density`, provide a nonparametric density estimate of the claims on both the original and logarithmic scale over the range of the data. Use this display to verify that the display is more interpretable on the logarithmic scale.
- b. Fit a normal distribution to logarithmic claims and compare the fitted distribution to the nonparametric (empirical) distribution. Interpret this comparison to mean that the lognormal distribution is an excellent candidate to represent these data.
- c. As an alternative, fit a Pareto distribution to the claims data using maximum likelihood. To check your work, do this in two ways. A basic approach is to create a log likelihood function and minimize it (using the function `optim`). A second approach is to the the `vglm` function from the `VGAM` package.
- d. We have fit  $X$  to be a Pareto distribution but wish to plot  $Y = \ln(X)$ . From Section 4.3.1.3, we saw that  $F_Y(y) = F_X(e^y)$  and  $f_Y(y) = e^y f_X(e^y)$ . Use this transformation to augment the plot in part (b) to include the Pareto distribution.

From this analysis, you learn that the lognormal and Pareto distribution fit the data approximately the same with the lognormal as a slight favorite.

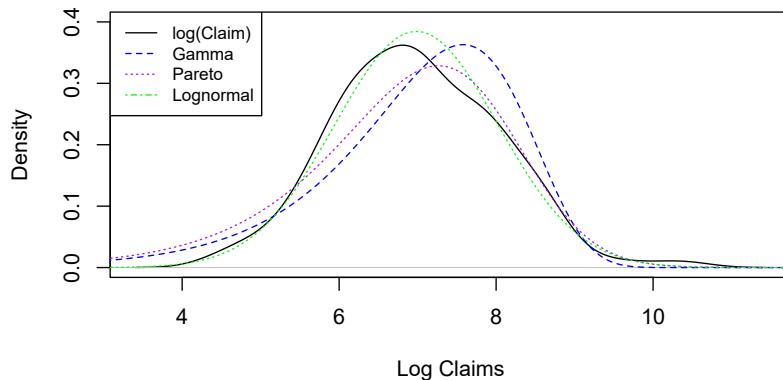
**Exercise 4.2. Wisconsin Property Fund.** Replicate the real-data example introduced in [Example 4.4.12](#) using the techniques demonstrated in Exercise 4.1.

**Exercise 4.3. Group Personal Accident.** This exercise is based on the data set introduced in [Exercise 1.2](#). We use incurred claims for all available years, still omitting those less than 10.

One can fit a distribution to the losses. An analysis, summarized in Figure 4.10,

shows the results from fitting via maximum likelihood the gamma, Pareto, and lognormal distributions to incurred losses. This figure suggests that the lognormal distribution appears to be the best fit.

Following the outlines in Exercises 4.1 and 4.2, fit these data via maximum likelihood and reproduce the figure that summarizes the results.



**FIGURE 4.10: Distribution of Group Personal Accident Losses with Superimposed Fitted Distributions**

## 4.6 Further Resources and Contributors

### Contributors

- **Zeinab Amin**, The American University in Cairo, is the principal author of the initial version and also the second edition of this chapter. **Edward (Jed) Frees** and **Lisa Gao**, University of Wisconsin-Madison, are the principal authors of the sections on nonparametric estimation which appeared in chapter 4 of the first edition of the text. Email: [zeinabha@aucegypt.edu](mailto:zeinabha@aucegypt.edu) for chapter comments and suggested improvements.
- Many helpful comments have been provided by Hirokazu (Iwahiro) Iwasawa, [iwahiro@bb.mbn.or.jp](mailto:iwahiro@bb.mbn.or.jp).
- Other chapter reviewers include: Rob Erhardt, Samuel Kolins, Tatjana Miljkovic, Michelle Xia, and Jorge Yslas.

**Further Readings and References**

Notable contributions include: Cummins and Derrig (2012), Frees and Valdez (2008), Klugman et al. (2012), Kreer et al. (2015), McDonald (1984), McDonald and Xu (1995), Tevet (2016), and Venter (1983).

If you would like additional practice with R coding, please visit our companion [LDA Short Course](#). In particular, see the [Modeling Loss Severity Chapter](#).

# 5

---

## *Modeling Claim Severity*

---

*Chapter Preview.* In Chapter 4 we explored the use of continuous as well as mixture distributions to model the random size of loss. Often the risk of loss is shared between the policyholder (the insured) and the insurer. Sharing risk can take the form of a deductible that is paid out-of-pocket of the insured before the insurer contributes to the loss, the form of a limit that caps the insured's liability for loss to a certain amount, or the form of a portion of the loss the insurer is responsible for covering after the insured covers his/her share of the cost, among other forms of cost-sharing. In Sections 5.1.1 to 5.1.3 we introduce the policy deductible feature of the insurance contract, the limited policy, and the co-insurance cost-sharing arrangement. In Section 5.1.4 we explore how insurance companies transfer part of the underlying insured risk by securing coverage from a reinsurer. Section 5.2 covers parametric estimation methods for modified data including grouped, censored and truncated data. In Section 5.3 we apply some non-parametric estimation tools like the ogive estimator, the plug-in principle, the Kaplan-Meier product-limit estimator, and the Nelson Aalon estimator on the modified data.

---

### **5.1 Coverage Modifications**

---

In this section, you learn how to:

- Describe the policy deductible feature of the insurance contract, the limited policy, and the coinsurance factor.
- Describe the distinction between the loss incurred to the insured and the amount of paid claim by the insurer under different policy modifications.
- Derive the distribution functions and raw moments for the amount of paid claim by the insurer for the different insurance contracts.
- Calculate the percentage decrease in the expected payment of the insurer as a result of imposing the deductible.
- Describe the insurance mechanism for insurance companies (reinsurance).

- Calculate the raw moments of the amount retained by the primary insurer in the reinsurance agreement.
- 

In this section we evaluate the impacts of coverage modifications: a) deductibles, b) policy limit, c) coinsurance and d) inflation on insurer's costs.

### 5.1.1 Policy Deductibles

Under an ordinary deductible policy, the insured (policyholder) agrees to cover a fixed amount of an insurance claim before the insurer starts to pay. This fixed expense paid out of pocket is called the deductible and often denoted by  $d$ . If the loss exceeds  $d$  then the insurer is responsible for covering the loss  $X$  less the deductible  $d$ . Depending on the agreement, the deductible may apply to each covered loss or to the total losses during a defined benefit period (such as a month, year, etc.)

Deductibles reduce premiums for the policyholders by eliminating a large number of small claims, the costs associated with handling these claims, and the potential moral hazard arising from having insurance. Moral hazard occurs when the insured takes more risks, increasing the chances of loss due to perils insured against, knowing that the insurer will incur the cost (e.g. a policyholder with collision insurance may be encouraged to drive recklessly). The larger the deductible, the less the insured pays in premiums for an insurance policy.

Let  $X$  denote the loss incurred to the insured and  $Y$  denote the amount of paid claim by the insurer. Speaking of the benefit paid to the policyholder, we differentiate between two variables: The payment per loss and the payment per payment. The payment per loss variable, denoted by  $Y^L$  or  $(X - d)_+$  is left censored because values of  $X$  that are less than  $d$  are set equal to zero. This variable is defined as

$$Y^L = (X - d)_+ = \begin{cases} 0 & X \leq d, \\ X - d & X > d \end{cases}.$$

$Y^L$  is often referred to as left censored and shifted variable because the values below  $d$  are not ignored and all losses are shifted by a value  $d$ .

On the other hand, the payment per payment variable, denoted by  $Y^P$ , is defined only when there is a payment. Specifically,  $Y^P$  equals  $X - d$  on the event  $\{X > d\}$ , denoted as  $Y^P = X - d | X > d$ . Another way of expressing this that is commonly used is

$$Y^P = \begin{cases} \text{Undefined} & X \leq d \\ X - d & X > d. \end{cases}$$

Here,  $Y^P$  is often referred to as left truncated and shifted variable or excess loss variable because the claims smaller than  $d$  are not reported and values above  $d$  are shifted by  $d$ .

Even when the distribution of  $X$  is continuous, the distribution of  $Y^L$  is a hybrid combination of discrete and continuous components. The discrete part of the distribution is concentrated at  $Y = 0$  (when  $X \leq d$ ) and the continuous part is spread over the interval  $Y > 0$  (when  $X > d$ ). For the discrete part, the probability that no payment is made is the probability that losses fall below the deductible; that is,

$$\Pr(Y^L = 0) = \Pr(X \leq d) = F_X(d).$$

Using the transformation  $Y^L = X - d$  for the continuous part of the distribution, we can find the *pdf* of  $Y^L$  given by

$$f_{Y^L}(y) = \begin{cases} F_X(d) & y = 0 \\ f_X(y + d) & y > 0. \end{cases}$$

We can see that the payment per payment variable is the payment per loss variable conditional on the loss exceeding the deductible ( $X > d$ ); that is,  $Y^P = Y^L | X > d$ . Alternatively, it can be expressed as  $Y^P = (X - d) | X > d$ , that is,  $Y^P$  is the loss in excess of the deductible given that the loss exceeds the deductible. Hence, the *pdf* of  $Y^P$  is given by

$$f_{Y^P}(y) = \frac{f_X(y + d)}{1 - F_X(d)},$$

for  $y > 0$ . Accordingly, the distribution functions of  $Y^L$  and  $Y^P$  are given by

$$F_{Y^L}(y) = \begin{cases} F_X(d) & y = 0 \\ F_X(y + d) & y > 0, \end{cases}$$

and

$$F_{Y^P}(y) = \frac{F_X(y + d) - F_X(d)}{1 - F_X(d)},$$

for  $y > 0$ , respectively.

The raw moments of  $Y^L$  and  $Y^P$  can be found directly using the *pdf* of  $X$  as follows

$$E[(Y^L)^k] = \int_d^\infty (x - d)^k f_X(x) dx,$$

and

$$E[(Y^P)^k] = \frac{\int_d^\infty (x - d)^k f_X(x) dx}{1 - F_X(d)} = \frac{E[(Y^L)^k]}{1 - F_X(d)},$$

respectively. For  $k = 1$ , we can use the survival function to calculate  $E(Y^L)$  as

$$E(Y^L) = \int_d^\infty [1 - F_X(x)] dx.$$

This could be easily proved if we start with the initial definition of  $E(Y^L)$  and use integration by parts.

We have seen that the deductible  $d$  imposed on an insurance policy is the amount of loss that has to be paid out of pocket before the insurer makes any payment. The deductible  $d$  imposed on an insurance policy reduces the insurer's payment. The loss elimination ratio (LER) is the percentage decrease in the expected payment of the insurer as a result of imposing the deductible. It is defined as

$$LER = \frac{E(X) - E(Y^L)}{E(X)}.$$

A little less common type of policy deductible is the franchise deductible. The franchise deductible will apply to the policy in the same way as ordinary deductible except that when the loss exceeds the deductible  $d$ , the full loss is covered by the insurer. The payment per loss and payment per payment variables are defined as

$$Y^L = \begin{cases} 0 & X \leq d, \\ X & X > d, \end{cases}$$

and

$$Y^P = \begin{cases} \text{Undefined} & X \leq d, \\ X & X > d, \end{cases}$$

respectively.

**Example 5.1.1. Actuarial Exam Question.** A claim severity distribution is exponential with mean 1000. An insurance company will pay the amount of each claim in excess of a deductible of 100. Calculate the variance of the amount paid by the insurance company for one claim, including the possibility that the amount paid is 0.

**Example Solution.** Let  $Y^L$  denote the amount paid by the insurance company for one claim.

$$Y^L = (X - 100)_+ = \begin{cases} 0 & X \leq 100, \\ X - 100 & X > 100. \end{cases}$$

The first and second moments of  $Y^L$  are

$$E(Y^L) = \int_{100}^{\infty} (x - 100) f_X(x) dx = \int_{100}^{\infty} S_X(x) dx = 1000e^{-\frac{100}{1000}},$$

and

$$E[(Y^L)^2] = \int_{100}^{\infty} (x - 100)^2 f_X(x) dx = 2 \times 1000^2 e^{-\frac{100}{1000}}.$$

So,

$$\text{Var}(Y^L) = \left(2 \times 1000^2 e^{-\frac{100}{1000}}\right) - \left(1000 e^{-\frac{100}{1000}}\right)^2 = 990,944.$$

An arguably simpler path to the solution is to make use of the relationship between  $X$  and  $Y^P$ . If  $X$  is exponentially distributed with mean 1000, then  $Y^P$  is also exponentially distributed with the same mean, because of the memoryless property of the exponential distribution. Hence,  $E(Y^P) = 1000$  and

$$E[(Y^P)^2] = 2 \times 1000^2.$$

Using the relationship between  $Y^L$  and  $Y^P$  we find

$$E(Y^L) = E(Y^P) S_X(100) = 1000 e^{-\frac{100}{1000}}$$

$$E[(Y^L)^2] = E[(Y^P)^2] S_X(100) = 2 \times 1000^2 e^{-\frac{100}{1000}}.$$

The relationship between  $X$  and  $Y^P$  can also be used when dealing with the uniform or the Pareto distributions. You can easily show that if  $X$  is uniform over the interval  $(0, \theta)$  then  $Y^P$  is uniform over the interval  $(0, \theta - d)$  and if  $X$  is Pareto with parameters  $\alpha$  and  $\theta$  then  $Y^P$  is Pareto with parameters  $\alpha$  and  $\theta + d$ .

**Example 5.1.2. Actuarial Exam Question.** For an insurance:

- Losses have a density function

$$f_X(x) = \begin{cases} 0.02x & 0 < x < 10, \\ 0 & \text{elsewhere.} \end{cases}$$

- The insurance has an ordinary deductible of 4 per loss.
- $Y^P$  is the claim payment per payment random variable.

Calculate  $E(Y^P)$ .

**Example Solution.** We define  $Y^P$  as follows

$$Y^P = \begin{cases} \text{Undefined} & X \leq 4, \\ X - 4 & X > 4. \end{cases}$$

$$\text{So, } E(Y^P) = \int_4^{10} (x - 4) 0.02x dx / 1 - F_X(4) = \frac{2.88}{0.84} = 3.43.$$

Note that we divide by  $S_X(4) = 1 - F_X(4)$ , as this is the probability where the variable  $Y^P$  is defined.

**Example 5.1.3. Actuarial Exam Question.** You are given:

- Losses follow an exponential distribution with the same mean in all years.
- The loss elimination ratio this year is 70%.
- The ordinary deductible for the coming year is  $4/3$  of the current deductible.

Compute the loss elimination ratio for the coming year.

**Example Solution.** Let the losses  $X \sim Exp(\theta)$  and the deductible for the coming year  $d' = \frac{4}{3}d$ , the deductible of the current year. The *LER* for the current year is

$$\frac{E(X) - E(Y^L)}{E(X)} = \frac{\theta - \theta e^{-d/\theta}}{\theta} = 1 - e^{-d/\theta} = 0.7.$$

Then,  $e^{-d/\theta} = 0.3$ .

The *LER* for the coming year is

$$\begin{aligned} \frac{\theta - \theta \exp(-\frac{d'}{\theta})}{\theta} &= \frac{\theta - \theta \exp\left(-\frac{\frac{4}{3}d}{\theta}\right)}{\theta} \\ &= 1 - \exp\left(-\frac{\frac{4}{3}d}{\theta}\right) = 1 - (e^{-d/\theta})^{4/3} = 1 - 0.3^{4/3} = 0.8. \end{aligned}$$

### 5.1.2 Policy Limits

Under a limited policy, the insurer is responsible for covering the actual loss  $X$  up to the limit of its coverage. This fixed limit of coverage is called the policy limit and often denoted by  $u$ . If the loss exceeds the policy limit, the difference  $X - u$  has to be paid by the policyholder. While a higher policy limit means a higher payout to the insured, it is associated with a higher premium.

Let  $X$  denote the loss incurred to the insured and  $Y$  denote the amount of paid claim by the insurer. The variable  $Y$  is known as the *limited loss variable* and is denoted by  $X \wedge u$ . It is a right censored variable because values above  $u$  are set equal to  $u$ . The limited loss random variable  $Y$  is defined as

$$Y = X \wedge u = \begin{cases} X & X \leq u \\ u & X > u. \end{cases}$$

It can be seen that the distinction between  $Y^L$  and  $Y^P$  is not needed under limited policy as the insurer will always make a payment.

Using the definitions of  $(X - u)_+$  and  $(X \wedge u)$ , it can be easily seen that the expected payment without any coverage modification,  $X$ , is equal to the sum of the expected payments with deductible  $u$  and limit  $u$ . That is,  $X = (X - u)_+ + (X \wedge u)$ .

Even when the distribution of  $X$  is continuous, the distribution of  $Y$  is a hybrid combination of discrete and continuous components. The discrete part of the distribution is concentrated at  $Y = u$  (when  $X > u$ ), while the continuous part is spread over the interval  $Y < u$  (when  $X \leq u$ ). For the discrete part, the probability that the benefit paid is  $u$ , is the probability that the loss exceeds the policy limit  $u$ ; that is,

$$\Pr(Y = u) = \Pr(X > u) = 1 - F_X(u).$$

For the continuous part of the distribution  $Y = X$ , hence the *pdf* of  $Y$  is given by

$$f_Y(y) = \begin{cases} f_X(y) & 0 < y < u \\ 1 - F_X(u) & y = u. \end{cases}$$

Accordingly, the distribution function of  $Y$  is given by

$$F_Y(y) = \begin{cases} F_X(x) & 0 < y < u \\ 1 & y \geq u. \end{cases}$$

The raw moments of  $Y$  can be found directly using the *pdf* of  $X$  as follows

$$E(Y^k) = E[(X \wedge u)^k] = \int_0^u x^k f_X(x) dx + \int_u^\infty u^k f_X(x) dx = \int_0^u x^k f_X(x) dx + u^k [1 - F_X(u)].$$

An alternative expression using the survival function is

$$E[(X \wedge u)^k] = \int_0^u kx^{k-1} [1 - F_X(x)] dx.$$

In particular, for  $k = 1$ , this is

$$E(Y) = E(X \wedge u) = \int_0^u [1 - F_X(x)] dx.$$

This could be easily proved if we start with the initial definition of  $E(Y)$  and use integration by parts. Alternatively, see the following justification of this limited expectation result.

$$\begin{aligned} E[(X \wedge u)^k] &= E\left[\int_0^{X \wedge u} kx^{k-1} dx\right] \\ &= E\left[\int_0^u kx^{k-1} I(X > x) dx\right] \\ &= \int_0^u kx^{k-1} EI(X > x) dx \\ &= \int_0^u kx^{k-1} [1 - F_X(x)] dx. \end{aligned}$$

This approach uses the Fubini-Tonelli theorem to exchange the expectation and integration. Note that it does not make any continuity assumptions about the distribution of  $X$ .

When a loss is subject to a deductible  $d$  and a limit  $u$ , the per-loss variable  $Y^L$  is defined as

$$Y^L = \begin{cases} 0 & X \leq d \\ X - d & d < X \leq u \\ u - d & X > u. \end{cases}$$

Hence,  $Y^L$  can be expressed as  $Y^L = (X \wedge u) - (X \wedge d)$ .

**Example 5.1.4. Actuarial Exam Question.** Under a group insurance policy, an insurer agrees to pay 100% of the medical bills incurred during the year by employees of a small company, up to a maximum total of one million dollars. The total amount of bills incurred,  $X$ , has pdf

$$f_X(x) = \begin{cases} \frac{x(4-x)}{9} & 0 < x < 3 \\ 0 & \text{elsewhere.} \end{cases}$$

where  $x$  is measured in millions. Calculate the total amount, in millions of dollars, the insurer would expect to pay under this policy.

**Example Solution.** Define the total amount of bills paid by the insurer as

$$Y = X \wedge 1 = \begin{cases} X & X \leq 1 \\ 1 & X > 1. \end{cases}$$

$$\text{So } E(Y) = E(X \wedge 1) = \int_0^1 (x^2(4-x))/9 \, dx + 1 \cdot \int_1^3 (x(4-x))/9 \, dx = 0.935.$$

### 5.1.3 Coinsurance and Inflation

As we have seen in Section 5.1.1, the amount of loss retained by the policy-holder can be losses up to the deductible  $d$ . The retained loss can also be a percentage of the claim. The percentage  $\alpha$ , often referred to as the coinsurance factor, is the percentage of claim the insurance company is required to cover. If the policy is subject to an ordinary deductible and policy limit, coinsurance refers to the percentage of claim the insurer is required to cover, after imposing the ordinary deductible and policy limit. The payment per loss variable,  $Y^L$ , is defined as

$$Y^L = \begin{cases} 0 & X \leq d, \\ \alpha(X - d) & d < X \leq u, \\ \alpha(u - d) & X > u. \end{cases}$$

The maximum amount paid by the insurer in this case is  $\alpha(u - d)$ , while  $u$  is the maximum covered loss.

We have seen in Section 5.1.2 that when a loss is subject to both a deductible  $d$  and a limit  $u$  the per-loss variable  $Y^L$  can be expressed as  $Y^L = (X \wedge u) - (X \wedge d)$ . With coinsurance, this becomes  $Y^L$  can be expressed as  $Y^L = \alpha[(X \wedge u) - (X \wedge d)]$ .

The  $k$ -th raw moment of  $Y^L$  is given by

$$\mathbb{E}[(Y^L)^k] = \int_d^u [\alpha(x - d)]^k f_X(x) dx + [\alpha(u - d)]^k [1 - F_X(u)].$$

A growth factor  $(1 + r)$  may be applied to  $X$  resulting in an inflated loss random variable  $(1 + r)X$  (the prespecified  $d$  and  $u$  remain unchanged). The resulting per loss variable can be written as

$$Y^L = \begin{cases} 0 & X \leq \frac{d}{1+r} \\ \alpha[(1+r)X - d] & \frac{d}{1+r} < X \leq \frac{u}{1+r} \\ \alpha(u - d) & X > \frac{u}{1+r}. \end{cases}$$

The first and second moments of  $Y^L$  can be expressed as

$$\mathbb{E}(Y^L) = \alpha(1+r) \left[ \mathbb{E}\left(X \wedge \frac{u}{1+r}\right) - \mathbb{E}\left(X \wedge \frac{d}{1+r}\right) \right],$$

and

$$\begin{aligned} \mathbb{E}[(Y^L)^2] &= \alpha^2(1+r)^2 \left\{ \mathbb{E}\left[\left(X \wedge \frac{u}{1+r}\right)^2\right] - \mathbb{E}\left[\left(X \wedge \frac{d}{1+r}\right)^2\right] \right. \\ &\quad \left. - 2\left(\frac{d}{1+r}\right) [\mathbb{E}\left(X \wedge \frac{u}{1+r}\right) - \mathbb{E}\left(X \wedge \frac{d}{1+r}\right)] \right\}, \end{aligned}$$

respectively.

The formulas given for the first and second moments of  $Y^L$  are general. Under full coverage,  $\alpha = 1$ ,  $r = 0$ ,  $u = \infty$ ,  $d = 0$  and  $\mathbb{E}(Y^L)$  reduces to  $\mathbb{E}(X)$ . If only an ordinary deductible is imposed,  $\alpha = 1$ ,  $r = 0$ ,  $u = \infty$  and  $\mathbb{E}(Y^L)$  reduces to  $\mathbb{E}(X) - \mathbb{E}(X \wedge d)$ . If only a policy limit is imposed  $\alpha = 1$ ,  $r = 0$ ,  $d = 0$  and  $\mathbb{E}(Y^L)$  reduces to  $\mathbb{E}(X \wedge u)$ .

**Example 5.1.5. Actuarial Exam Question.** The ground up loss random variable for a health insurance policy in 2006 is modeled with  $X$ , a random variable with an exponential distribution having mean 1000. An insurance policy pays the loss above an ordinary deductible of 100, with a maximum annual payment of 500. The ground up loss random variable is expected to be 5% larger in 2007, but the insurance in 2007 has the same deductible and maximum payment as in 2006. Find the percentage increase in the expected cost per payment from 2006 to 2007.

**Example Solution.** We define the amount per loss  $Y^L$  in both years as

$$Y_{2006}^L = \begin{cases} 0 & X \leq 100, \\ X - 100 & 100 < X \leq 600, \\ 500 & X > 600. \end{cases}$$

$$Y_{2007}^L = \begin{cases} 0 & X \leq 95.24, \\ 1.05X - 100 & 95.24 < X \leq 571.43, \\ 500 & X > 571.43. \end{cases}$$

So,

$$\begin{aligned} E(Y_{2006}^L) &= E(X \wedge 600) - E(X \wedge 100) \\ &= 1000 \left(1 - e^{-\frac{600}{1000}}\right) - 1000 \left(1 - e^{-\frac{100}{1000}}\right) \\ &= 356.026. \end{aligned}$$

Further,

$$\begin{aligned} E(Y_{2007}^L) &= 1.05 [E(X \wedge 571.43) - E(X \wedge 95.24)] \\ &= 1.05 \left[1000 \left(1 - e^{-\frac{571.43}{1000}}\right) - 1000 \left(1 - e^{-\frac{95.24}{1000}}\right)\right] \\ &= 361.659. \end{aligned}$$

$$\begin{aligned} E(Y_{2006}^P) &= \frac{356.026}{e^{-(100/1000)}} = 393.469. \\ E(Y_{2007}^P) &= \frac{361.659}{e^{-(95.24/1000)}} = 397.797. \end{aligned}$$

Because  $\frac{E(Y_{2007}^P)}{E(Y_{2006}^P)} - 1 = 0.011$ , there is an increase of 1.1 percent from 2006 to 2007. Due to the policy limit, the cost per payment event grew by only 1.1 percent between 2006 and 2007 even though the ground up losses increased by 5 percent between the two years.

### 5.1.4 Reinsurance

In Section 5.1.1 we introduced the policy deductible feature of the insurance contract. In this feature, there is a contractual arrangement under which an insured transfers part of the risk by securing coverage from an insurer in return for an insurance premium. Under that policy, the insured must pay all losses up to the deductible, and the insurer only pays the amount (if any) above the deductible. We now introduce reinsurance, a mechanism of insurance for insurance companies. Reinsurance is a contractual arrangement under which an insurer transfers part of the underlying insured risk by securing coverage from another insurer (referred to as a reinsurer) in return for a reinsurance premium. Although reinsurance involves a relationship between three parties: the original insured, the insurer (often referred to as cedent or cedant) and the reinsurer, the parties of the reinsurance agreement are only the primary insurer and the reinsurer. There is no contractual agreement between the

original insured and the reinsurer. Though many different types of reinsurance contracts exist, a common form is excess of loss coverage. In such contracts, the primary insurer must make all required payments to the insured until the primary insurer's total payments reach a fixed reinsurance deductible. The reinsurer is then only responsible for paying losses above the reinsurance deductible. The maximum amount retained by the primary insurer in the reinsurance agreement (the reinsurance deductible) is called retention.

Reinsurance arrangements allow insurers with limited financial resources to increase the capacity to write insurance and meet client requests for larger insurance coverage while reducing the impact of potential losses and protecting the insurance company against catastrophic losses. Reinsurance also allows the primary insurer to benefit from underwriting skills, expertise and proficient complex claim file handling of the larger reinsurance companies.

**Example 5.1.6. Actuarial Exam Question.** Losses arising in a certain portfolio have a two-parameter Pareto distribution with  $\alpha = 5$  and  $\theta = 3,600$ . A reinsurance arrangement has been made, under which (a) the reinsurer accepts 15% of losses up to  $u = 5,000$  and all amounts in excess of 5,000 and (b) the insurer pays for the remaining losses.

- Express the random variables for the reinsurer's and the insurer's payments as a function of  $X$ , the portfolio losses.
- Calculate the mean amount paid on a single claim by the insurer.
- By assuming that the upper limit is  $u = \infty$ , calculate an upper bound on the standard deviation of the amount paid on a single claim by the insurer (retaining the 15% copayment).

**Example Solution.**

- a). The reinsurer's portion is

$$Y_{reinsurer} = \begin{cases} 0.15X & X < 5000, \\ 0.15(5000) + X - 5000 & X \geq 5000 \end{cases}.$$

and the insurer's portion is

$$Y_{insurer} = \begin{cases} 0.85X & X < 5000, \\ 0.85(5000) & X \geq 5000 \end{cases} = 0.85(X \wedge 5000).$$

b) Using the limited expected value tables for the Pareto distribution, we have

$$\begin{aligned} E(Y_{insurer}) &= 0.85 E(X \wedge 5000) = 0.85 \frac{\theta}{\alpha-1} \left[ 1 - \left( \frac{\theta}{5000+\theta} \right)^{\alpha-1} \right] \\ &= 0.85 \frac{3600}{5-1} \left[ 1 - \left( \frac{3600}{5000+3600} \right)^{5-1} \right] = 741.5103. \end{aligned}$$

c) The unlimited variable is  $0.85X$ . For the first moment, we have

$$0.85 E X = 0.85 \frac{\theta}{\alpha-1} = 0.85 \frac{3600}{5-1} = 765.$$

For the second moment of the unlimited variable, we use the table of distributions to get

$$0.85^2 E X^2 = 0.85^2 \frac{\theta^2 \Gamma(2+1)\Gamma(\alpha-2)}{\Gamma(\alpha)} = 0.85^2 \frac{3600^2 \cdot 2 \cdot 2}{24} = 1560600.$$

Thus, the variance is  $1560600 - 765^2 = 975375$ . Alternatively, you can use the formula

$$0.85^2 \text{Var } X = 0.85^2 \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)} = 0.85^2 \frac{5(3600^2)}{(5-1)^2(5-2)} = 975375.$$

Taking square roots, the standard deviation is  $\sqrt{975375} \approx 987.6108$ .

---

Further discussions of reinsurance will be provided in Section ??.

---

## 5.2 Parametric Estimation using Modified Data

---

In this section, you learn how to:

- Describe grouped, censored, and truncated data
  - Estimate parametric distributions based on grouped, censored, and truncated data
- 

Basic theory and many applications are based on *individual* observations that are “*complete*” and “*unmodified*,” as we have seen in the Chapter 4. Section 5.1.1 introduced the concept of observations that are “*modified*” due to two common types of limitations: **censoring** and **truncation**. For example, it is common to think about an insurance deductible as producing data that are

truncated (from the left) or policy limits as yielding data that are censored (from the right). This viewpoint is from the primary insurer (the seller of the insurance). Another viewpoint is that of a reinsurer (an insurer of an insurance company) that will be discussed more in Chapter ???. A reinsurer may not observe a claim smaller than an amount, only that a claim exists; this is an example of censoring from the left. So, in this section, we cover the full gamut of alternatives. Specifically, this section will address parametric estimation methods for three alternatives to individual, complete, and unmodified data: **interval-censored** data available only in groups, data that are limited or **censored**, and data that may not be observed due to **truncation**.

### 5.2.1 Parametric Estimation using Grouped Data

Consider a sample of size  $n$  observed from the distribution  $F(\cdot)$ , but in groups so that we only know the group into which each observation fell, not the exact value. This is referred to as **grouped** or **interval-censored** data. For example, we may be looking at two successive years of annual employee records. People employed in the first year but not the second have left sometime during the year. With an exact departure date (individual data), we could compute the amount of time that they were with the firm. Without the departure date (grouped data), we only know that they departed sometime during a year-long interval.

Formalizing this idea, suppose there are  $k$  groups or intervals delimited by boundaries  $c_0 < c_1 < \dots < c_k$ . For each observation, we only observe the interval into which it fell (e.g.  $(c_{j-1}, c_j]$ ), not the exact value. Thus, we only know the number of observations in each interval. The constants  $\{c_0 < c_1 < \dots < c_k\}$  form some partition of the domain of  $F(\cdot)$ . Then the probability of an observation  $X_i$  falling in the  $j$ th interval is

$$\Pr(X_i \in (c_{j-1}, c_j]) = F(c_j) - F(c_{j-1}).$$

The corresponding probability mass function for an observation is

$$\begin{aligned} f(x) &= \begin{cases} F(c_1) - F(c_0) & \text{if } x \in (c_0, c_1] \\ \vdots & \vdots \\ F(c_k) - F(c_{k-1}) & \text{if } x \in (c_{k-1}, c_k] \end{cases} \\ &= \prod_{j=1}^k \{F(c_j) - F(c_{j-1})\}^{I(x \in (c_{j-1}, c_j])} \end{aligned}$$

Now, define  $n_j$  to be the number of observations that fall in the  $j$ th interval,  $(c_{j-1}, c_j]$ . Thus, the likelihood function (with respect to the parameter(s)  $\theta$ )

is

$$L(\theta) = \prod_{j=1}^n f(x_i) = \prod_{j=1}^k \{F(c_j) - F(c_{j-1})\}^{n_j}$$

And the log-likelihood function is

$$l(\theta) = \log L(\theta) = \log \prod_{j=1}^n f(x_i) = \sum_{j=1}^k n_j \log \{F(c_j) - F(c_{j-1})\}$$

Maximizing the likelihood function (or equivalently, maximizing the log-likelihood function) would then produce the maximum likelihood estimates for grouped data.

**Example 5.2.1. Actuarial Exam Question.** You are given:

- (i) Losses follow an exponential distribution with mean  $\theta$ .
- (ii) A random sample of 20 losses is distributed as follows:

Loss Range	Frequency
[0, 1000]	7
(1000, 2000]	6
(2000, $\infty$ )	7

Calculate the maximum likelihood estimate of  $\theta$ .

**Example Solution.**

$$\begin{aligned} L(\theta) &= F(1000)^7 [F(2000) - F(1000)]^6 [1 - F(2000)]^7 \\ &= (1 - e^{-1000/\theta})^7 (e^{-1000/\theta} - e^{-2000/\theta})^6 (e^{-2000/\theta})^7 \\ &= (1 - p)^7 (p - p^2)^6 (p^2)^7 \\ &= p^{20} (1 - p)^{13} \end{aligned}$$

where  $p = e^{-1000/\theta}$ . Maximizing this expression with respect to  $p$  is equivalent to maximizing the likelihood with respect to  $\theta$ . The maximum occurs at  $p = \frac{20}{33}$  and so  $\hat{\theta} = \frac{-1000}{\log(20/33)} = 1996.90$ .

### 5.2.2 Censored Data

**Censoring** occurs when we record only a limited value of an observation. The most common form is **right-censoring**, in which we record the *smaller* of the “true” dependent variable and a censoring value. Using notation, let  $X$  represent an outcome of interest, such as the loss due to an insured event or

time until an event. Let  $C_U$  denote the censoring amount. With right-censored observations, we record  $X_U^* = \min(X, C_U) = X \wedge C_U$ . We also record whether or not censoring has occurred. Let  $\delta_U = I(X \leq C_U)$  be a binary variable that is 0 if censoring occurs and 1 if it does not, that is,  $\delta_U$  indicates whether or not  $X$  is uncensored.

For an example that we saw in Section 5.1.2,  $C_U$  may represent the upper limit of coverage of an insurance policy (we used  $u$  for the upper limit in that section). The loss may exceed the amount  $C_U$ , but the insurer only has  $C_U$  in its records as the amount paid out and does not have the amount of the actual loss  $X$  in its records.

Similarly, with **left-censoring**, we record the *larger* of a variable of interest and a censoring variable. If  $C_L$  is used to represent the censoring amount, we record  $X_L^* = \max(X, C_L)$  along with the censoring indicator  $\delta_L = I(X > C_L)$ .

As an example, we gave a brief introduction to reinsurance (insurance for insurers) in Section 5.1.4 and more is given in Chapter ???. Suppose a reinsurer will cover insurer losses greater than  $C_L$ ; this means that the reinsurer is responsible for the excess of  $X_L^*$  over  $C_L$ . Using notation, the loss of the reinsurer is  $Y = X_L^* - C_L$ . To see this, first consider the case where the policyholder loss  $X < C_L$ . Then, the insurer will pay the entire claim and  $Y = C_L - C_L = 0$ , no loss for the reinsurer. For contrast, if the loss  $X \geq C_L$ , then  $Y = X - C_L$  represents the reinsurer's retained claims. Put another way, if a loss occurs, the reinsurer records the actual amount if it exceeds the limit  $C_L$  and otherwise it only records that it had a loss of 0.

### 5.2.3 Truncated Data

Censored observations are recorded for study, although in a limited form. In contrast, **truncated** outcomes are a type of missing data. An outcome is potentially truncated when the availability of an observation depends on the outcome.

In insurance, it is common for observations to be **left-truncated** at  $C_L$  when the amount is

$$Y = \begin{cases} \text{we do not observe } X & X \leq C_L \\ X & X > C_L \end{cases} .$$

In other words, if  $X$  is less than the threshold  $C_L$ , then it is not observed.

For an example we saw in Section 5.1.1,  $C_L$  may represent the deductible of an insurance policy (we used  $d$  for the deductible in that section). If the insured loss is less than the deductible, then the insurer may not observe or record the loss at all. If the loss exceeds the deductible, then the excess  $X - C_L$  is the

claim that the insurer covers. In Section 5.1.1, we defined the per payment loss to be

$$Y^P = \begin{cases} \text{Undefined} & X \leq d \\ X - d & X > d \end{cases},$$

so that if a loss exceeds a deductible, we record the excess amount  $X - d$ . This is very important when considering amounts that the insurer will pay. However, for estimation purposes of this section, it matters little if we subtract a known constant such as  $C_L = d$ . So, for our truncated variable  $Y$ , we use the simpler convention and do not subtract  $d$ .

Similarly for **right-truncated** data, if  $X$  exceeds a threshold  $C_U$ , then it is not observed. In this case, the amount is

$$Y = \begin{cases} X & X \leq C_U \\ \text{we do not observe } X & X > C_U. \end{cases}$$

Classic examples of truncation from the right include  $X$  as a measure of distance to a star. When the distance exceeds a certain level  $C_U$ , the star is no longer observable.

Figure 5.1 compares truncated and censored observations.

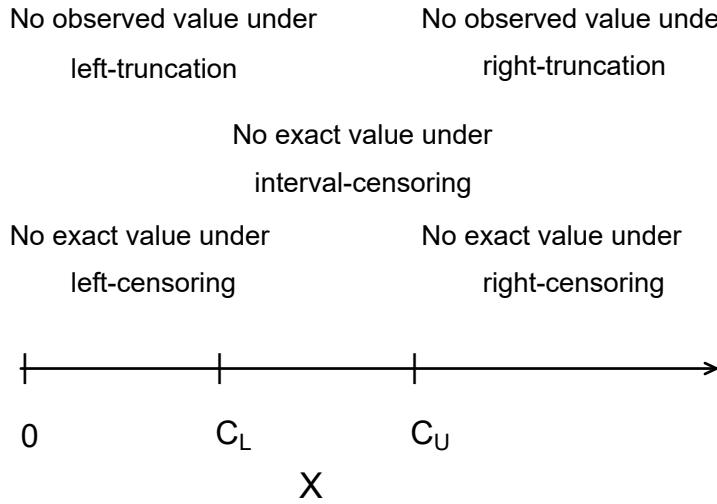


FIGURE 5.1: Censoring and Truncation

**Example – Mortality Study.** Suppose that you are conducting a two-year study of mortality of high-risk subjects, beginning January 1, 2010 and finishing January 1, 2012. Figure 5.2 graphically portrays the six types of subjects recruited. For each subject, the beginning of the arrow represents that the subject was recruited and the arrow end represents the event time where in this example the event represents death. The arrow represents exposure time.

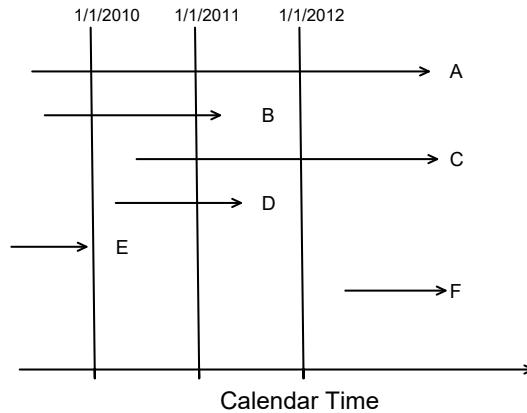


FIGURE 5.2: Timeline for Several Subjects on Test in a Mortality Study

- **Type A - Right-censored.** This subject is alive at the beginning and the end of the study. Because the time of death is not known by the end of the study, it is right-censored. Most subjects are Type A.
- **Type B - Complete** information is available for a type B subject. The subject is alive at the beginning of the study and the death occurs within the observation period.
- **Type C - Right-censored and left-truncated.** A type C subject is right-censored, in that death occurs after the observation period. However, the subject entered after the start of the study and is said to have a *delayed entry time*. Because the subject would not have been observed had death occurred before entry, it is left-truncated.
- **Type D - Left-truncated.** A type D subject also has delayed entry. Because death occurs within the observation period, this subject is not right censored.
- **Type E - Left-truncated.** A type E subject is not included in the study because death occurs prior to the observation period.
- **Type F - Right-truncated.** Similarly, a type F subject is not included because the entry time occurs after the observation period.

To summarize, for outcome  $X$  and constants  $C_L$  and  $C_U$ ,

Limitation Type	Limited Variable	Recording Information
right censoring	$X_U^* = \min(X, C_U)$	$\delta_U = I(X \leq C_U)$
left censoring	$X_L^* = \max(X, C_L)$	$\delta_L = I(X > C_L)$
interval censoring		
right truncation	$X$	observe $X$ if $X \leq C_U$
left truncation	$X$	observe $X$ if $X > C_L$

#### 5.2.4 Parametric Estimation using Censored and Truncated Data

For simplicity, we assume non-random censoring amounts and a continuous outcome  $X$ . To begin, consider the case of right-censored data where we record  $X_U^* = \min(X, C_U)$  and censoring indicator  $\delta = I(X \leq C_U)$ . If censoring occurs so that  $\delta = 0$ , then  $X > C_U$  and the likelihood is  $\Pr(X > C_U) = 1 - F(C_U)$ . If censoring does not occur so that  $\delta = 1$ , then  $X \leq C_U$  and the likelihood is  $f(x)$ . Summarizing, we have the likelihood of a single observation as

$$\begin{cases} 1 - F(C_U) & \text{if } \delta = 0 \\ f(x) & \text{if } \delta = 1 \end{cases} = \{f(x)\}^\delta \{1 - F(C_U)\}^{1-\delta}.$$

The right-hand expression allows us to present the likelihood more compactly. Now, for an *iid* sample of size  $n$ , the likelihood is

$$L(\theta) = \prod_{i=1}^n \{f(x_i)\}^{\delta_i} \{1 - F(C_{Ui})\}^{1-\delta_i} = \prod_{\delta_i=1} f(x_i) \prod_{\delta_i=0} \{1 - F(C_{Ui})\},$$

with potential censoring times  $\{C_{U1}, \dots, C_{Un}\}$ . Here, the notation “ $\prod_{\delta_i=1}$ ” means to take the product over uncensored observations, and similarly for “ $\prod_{\delta_i=0}$ ”.

On the other hand, truncated data are handled in likelihood inference via conditional probabilities. Specifically, we adjust the likelihood contribution by dividing by the probability that the variable was observed. To summarize, we have the following contributions to the likelihood function for six types of outcomes:

Outcome	Likelihood Contribution
exact value	$f(x)$
right-censoring	$1 - F(C_U)$
left-censoring	$F(C_L)$
right-truncation	$f(x)/F(C_U)$
left-truncation	$f(x)/(1 - F(C_L))$
interval-censoring	$F(C_U) - F(C_L)$

For known outcomes and censored data, the likelihood is

$$L(\theta) = \prod_E f(x_i) \prod_R \{1 - F(C_{Ui})\} \prod_L F(C_{Li}) \prod_I (F(C_{Ui}) - F(C_{Li})),$$

where “ $\prod_E$ ” is the product over observations with *Exact* values, and similarly for *Right-*, *Left-* and *Interval-censoring*.

For right-censored and left-truncated data, the likelihood is

$$L(\theta) = \prod_E \frac{f(x_i)}{1 - F(C_{Li})} \prod_R \frac{1 - F(C_{Ui})}{1 - F(C_{Li})},$$

and similarly for other combinations. To get further insights, consider the following.

**Special Case: Exponential Distribution.** Consider data that are right-censored and left-truncated, with random variables  $X_i$  that are exponentially distributed with mean  $\theta$ . With these specifications, recall that  $f(x) = \theta^{-1} \exp(-x/\theta)$  and  $F(x) = 1 - \exp(-x/\theta)$ .

For this special case, the log-likelihood is

$$\begin{aligned} l(\theta) &= \sum_E \{\log f(x_i) - \log(1 - F(C_{Li}))\} + \sum_R \{\log(1 - F(C_{Ui})) - \log(1 - F(C_{Li}))\} \\ &= \sum_E (-\log \theta - (x_i - C_{Li})/\theta) - \sum_R (C_{Ui} - C_{Li})/\theta. \end{aligned}$$

To simplify the notation, define  $\delta_i = I(X_i < C_{Ui})$  to be a binary variable that indicates right-censoring. Let  $X_i^{**} = \min(X_i, C_{Ui}) - C_{Li}$  be the amount that the observed variable exceeds the lower truncation limit. With this, the log-likelihood is

$$l(\theta) = - \sum_{i=1}^n \left( (1 - \delta_i) \log \theta + \frac{x_i^{**}}{\theta} \right) \quad (5.1)$$

Taking derivatives with respect to the parameter  $\theta$  and setting it equal to zero yields the maximum likelihood estimator

$$\hat{\theta} = \frac{1}{n_u} \sum_{i=1}^n x_i^{**},$$

where  $n_u = \sum_i (1 - \delta_i)$  is the number of uncensored observations.

**Example 5.2.2. Actuarial Exam Question.** You are given:

- (i) A sample of losses is: 600 700 900
- (ii) No information is available about losses of 500 or less.
- (iii) Losses are assumed to follow an exponential distribution with mean  $\theta$ .

Calculate the maximum likelihood estimate of  $\theta$ .

**Example Solution.** These observations are truncated at 500. The contribution of each observation to the likelihood function is

$$\frac{f(x)}{1 - F(500)} = \frac{\theta^{-1}e^{-x/\theta}}{e^{-500/\theta}}$$

Then the likelihood function is

$$L(\theta) = \frac{\theta^{-1}e^{-600/\theta}\theta^{-1}e^{-700/\theta}\theta^{-1}e^{-900/\theta}}{(e^{-500/\theta})^3} = \theta^{-3}e^{-700/\theta}$$

The log-likelihood is

$$l(\theta) = \log L(\theta) = -3\log\theta - 700\theta^{-1}$$

Maximizing this expression by setting the derivative with respect to  $\theta$  equal to 0, we have

$$L'(\theta) = -3\theta^{-1} + 700\theta^{-2} = 0 \Rightarrow \hat{\theta} = \frac{700}{3} = 233.33.$$

---

**Example 5.2.3. Actuarial Exam Question.** You are given the following information about a random sample:

- (i) The sample size equals five.
- (ii) The sample is from a Weibull distribution with  $\tau = 2$ .
- (iii) Two of the sample observations are known to exceed 50, and the remaining three observations are 20, 30, and 45.

Calculate the maximum likelihood estimate of  $\theta$ .

**Example Solution.** The likelihood function is

$$\begin{aligned} L(\theta) &= f(20)f(30)f(45)[1 - F(50)]^2 \\ &= \frac{2(20/\theta)^2 e^{-(20/\theta)^2}}{20} \frac{2(30/\theta)^2 e^{-(30/\theta)^2}}{30} \frac{2(45/\theta)^2 e^{-(45/\theta)^2}}{45} (e^{-(50/\theta)^2})^2 \\ &\propto \frac{1}{\theta^6} e^{-8325/\theta^2} \end{aligned}$$

The natural logarithm of the above expression is  $-6 \log \theta - \frac{8325}{\theta^2}$ . Maximizing this expression by setting its derivative to 0, we get

$$\frac{-6}{\theta} + \frac{16650}{\theta^3} = 0 \Rightarrow \hat{\theta} = \left( \frac{16650}{6} \right)^{\frac{1}{2}} = 52.6783.$$

## 5.3 Nonparametric Estimation using Modified Data

In this section, you learn how to:

- Estimate the distribution function for grouped data using the ogive.
- Create a nonparametric estimator of the loss elimination ratio using the plug-in principle.
- Apply the Kaplan-Meier product-limit and the Nelson Aalon estimators to estimate the distribution function in the presence of censoring.
- Apply Greenwood's formula to estimate the variance of the product-limit estimator.

Nonparametric estimators provide useful benchmarks, so it is helpful to understand the estimation procedures for grouped, censored, and truncated data.

### 5.3.1 Grouped Data

As we have seen in Section 5.2.1, observations may be grouped (also referred to as interval censored) in the sense that we only observe them as belonging in one of  $k$  intervals of the form  $(c_{j-1}, c_j]$ , for  $j = 1, \dots, k$ . At the boundaries, the empirical distribution function is defined in the usual way:

$$F_n(c_j) = \frac{\text{number of observations } \leq c_j}{n}.$$

**Ogive Estimator.** For other values of  $x \in (c_{j-1}, c_j)$ , we can estimate the distribution function with the ogive estimator, which linearly interpolates between  $F_n(c_{j-1})$  and  $F_n(c_j)$ , i.e. the values of the boundaries  $F_n(c_{j-1})$  and  $F_n(c_j)$  are connected with a straight line. This can formally be expressed as

$$F_n(x) = \frac{c_j - x}{c_j - c_{j-1}} F_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} F_n(c_j) \quad \text{for } c_{j-1} \leq x < c_j$$

The corresponding density is

$$f_n(x) = F'_n(x) = \frac{F_n(c_j) - F_n(c_{j-1})}{c_j - c_{j-1}} \quad \text{for } c_{j-1} < x < c_j.$$


---

**Example 5.3.1. Actuarial Exam Question.** You are given the following information regarding claim sizes for 100 claims:

Claim Size	Number of Claims
0 – 1,000	16
1,000 – 3,000	22
3,000 – 5,000	25
5,000 – 10,000	18
10,000 – 25,000	10
25,000 – 50,000	5
50,000 – 100,000	3
over 100,000	1

Using the ogive, calculate the estimate of the probability that a randomly chosen claim is between 2000 and 6000.

**Example Solution.** At the boundaries, the empirical distribution function is defined in the usual way, so we have

$$F_{100}(1000) = 0.16, \quad F_{100}(3000) = 0.38, \quad F_{100}(5000) = 0.63, \quad F_{100}(10000) = 0.81.$$

For other claim sizes, the ogive estimator linearly interpolates between these values:

$$\begin{aligned} F_{100}(2000) &= 0.5F_{100}(1000) + 0.5F_{100}(3000) = 0.5(0.16) + 0.5(0.38) = 0.27 \\ F_{100}(6000) &= 0.8F_{100}(5000) + 0.2F_{100}(10000) = 0.8(0.63) + 0.2(0.81) = 0.666 \end{aligned}$$

Thus, the probability that a claim is between 2000 and 6000 is  $F_{100}(6000) - F_{100}(2000) = 0.666 - 0.27 = 0.396$ .

### 5.3.2 Plug-in Principle

One way to create a nonparametric estimator of some quantity is to use the *analog* or plug-in principle where one replaces the unknown cdf  $F$  with a known estimate such as the empirical cdf  $F_n$ . So, if we are trying to estimate  $E[g(X)] = E_F[g(X)]$  for a generic function  $g$ , then we define a nonparametric estimator to be  $E_{F_n}[g(X)] = n^{-1} \sum_{i=1}^n g(X_i)$ .

To see how this works, as a special case of  $g$  we consider the loss per payment random variable is  $Y = (X - d)_+$  and the *loss elimination ratio* introduced in Section 4.4.1. We can express this as

$$LER(d) = \frac{E[X - (X - d)_+]}{E[X]} = \frac{E[\min(X, d)]}{E[X]},$$

for a fixed deductible  $d$ .

#### Example. 5.3.2. Bodily Injury Claims and Loss Elimination Ratios

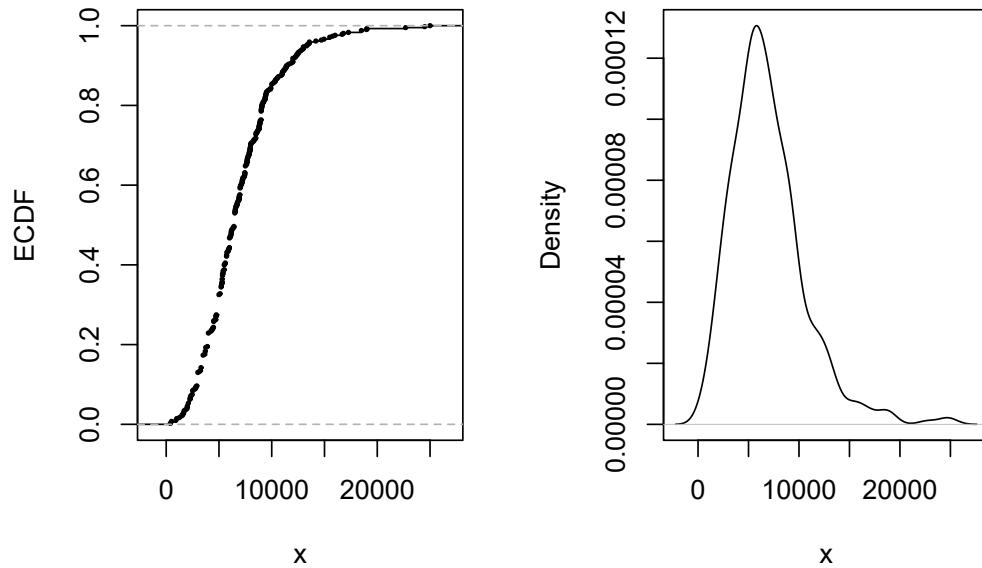
We use a sample of 432 closed auto claims from Boston from [Derrig et al. \(2001\)](#). Losses are recorded for payments due to bodily injuries in auto accidents. Losses are not subject to deductibles but are limited by various maximum coverage amounts that are also available in the data. It turns out that only 17 out of 432 ( $\approx 4\%$ ) were subject to these policy limits and so we ignore these data for this illustration.

The average loss paid is 6906 in U.S. dollars. Figure 5.3 shows other aspects of the distribution. Specifically, the left-hand panel shows the empirical distribution function, the right-hand panel gives a nonparametric density plot.

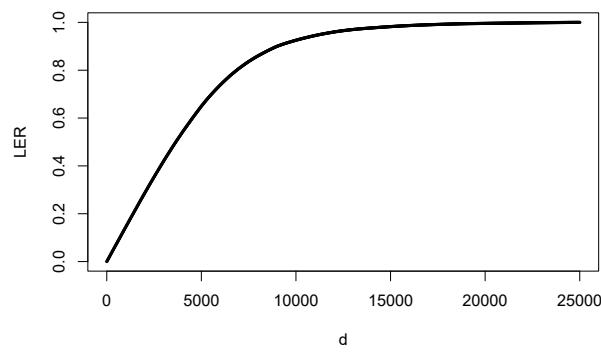
The impact of bodily injury losses can be mitigated by the imposition of limits or purchasing reinsurance policies (see Section 10.3). To quantify the impact of these risk mitigation tools, it is common to compute the *loss elimination ratio (LER)* as introduced in Section 4.4.1. The distribution function is not available and so must be estimated in some way. Using the plug-in principle, a nonparametric estimator can be defined as

$$LER_n(d) = \frac{n^{-1} \sum_{i=1}^n \min(X_i, d)}{n^{-1} \sum_{i=1}^n X_i} = \frac{\sum_{i=1}^n \min(X_i, d)}{\sum_{i=1}^n X_i}.$$

Figure 5.4 shows the estimator  $LER_n(d)$  for various choices of  $d$ . For example, at  $d = 1,000$ , we have  $LER_n(1000) \approx 0.1442$ . Thus, imposing a limit of 1,000 means that expected retained claims are 14.42 percent lower when compared to expected claims with a zero deductible.



**FIGURE 5.3: Bodily Injury Claims.** The left-hand panel gives the empirical distribution function. The right-hand panel presents a nonparametric density plot.



**FIGURE 5.4: LER for Bodily Injury Claims.** The figure presents the loss elimination ratio (LER) as a function of deductible  $d$ .

### 5.3.3 Right-Censored Empirical Distribution Function

It can be useful to calibrate parametric estimators with nonparametric methods that do not rely on a parametric form of the distribution. The product-limit estimator due to (Kaplan and Meier, 1958) is a well-known estimator of the distribution function in the presence of censoring.

**Motivation for the Kaplan-Meier Product Limit Estimator.** To explain why the product-limit works so well with censored observations, let us first return to the “usual” case without censoring. Here, the empirical distribution function  $F_n(x)$  is an *unbiased* estimator of the distribution function  $F(x)$ . This is because  $F_n(x)$  is the average of indicator variables each of which are unbiased, that is,  $E[I(X_i \leq x)] = \Pr(X_i \leq x) = F(x)$ .

Now suppose the random outcome is censored on the right by a limiting amount, say,  $C_U$ , so that we record the smaller of the two,  $X^* = \min(X, C_U)$ . For values of  $x$  that are smaller than  $C_U$ , the indicator variable still provides an unbiased estimator of the distribution function before we reach the censoring limit. That is,  $E[I(X^* \leq x)] = F(x)$  because  $I(X^* \leq x) = I(X \leq x)$  for  $x < C_U$ . In the same way,  $E[I(X^* > x)] = 1 - F(x) = S(x)$ . But, for  $x > C_U$ ,  $I(X^* \leq x)$  is in general not an unbiased estimator of  $F(x)$ .

As an alternative, consider *two* random variables that have different censoring limits. For illustration, suppose that we observe  $X_1^* = \min(X_1, 5)$  and  $X_2^* = \min(X_2, 10)$  where  $X_1$  and  $X_2$  are independent draws from the same distribution. For  $x \leq 5$ , the empirical distribution function  $F_2(x)$  is an unbiased estimator of  $F(x)$ . However, for  $5 < x \leq 10$ , the first observation cannot be used for the distribution function because of the censoring limitation. Instead, the strategy developed by (Kaplan and Meier, 1958) is to use  $S_2(5)$  as an estimator of  $S(5)$  and then to use the second observation to estimate the survival function conditional on survival to time 5,  $\Pr(X > x | X > 5) = \frac{S(x)}{S(5)}$ . Specifically, for  $5 < x \leq 10$ , the estimator of the survival function is

$$\hat{S}(x) = S_2(5) \times I(X_2^* > x).$$

**Kaplan-Meier Product Limit Estimator.** Extending this idea, for each observation  $i$ , let  $u_i$  be the upper censoring limit ( $= \infty$  if no censoring). Thus, the recorded value is  $x_i$  in the case of no censoring and  $u_i$  if there is censoring. Let  $t_1 < \dots < t_k$  be  $k$  distinct points at which an uncensored loss occurs, and let  $s_j$  be the number of uncensored losses  $x_i$ 's at  $t_j$ . The corresponding risk set is the number of observations that are active (not censored) at a value *less than*  $t_j$ , denoted as  $R_j = \sum_{i=1}^n I(x_i \geq t_j) + \sum_{i=1}^n I(u_i \geq t_j)$ .

With this notation, the **product-limit estimator** of the distribution function

is

$$\hat{F}(x) = \begin{cases} 0 & x < t_1 \\ 1 - \prod_{j:t_j \leq x} \left(1 - \frac{s_j}{R_j}\right) & x \geq t_1 \end{cases}. \quad (5.2)$$

For example, if  $x$  is smaller than the smallest uncensored loss, then  $x < t_1$  and  $\hat{F}(x) = 0$ . As another example, if  $x$  falls between then second and third smallest uncensored losses, then  $x \in (t_2, t_3]$  and  $\hat{F}(x) = 1 - \left(1 - \frac{s_1}{R_1}\right) \left(1 - \frac{s_2}{R_2}\right)$ .

As usual, the corresponding estimate of the survival function is  $\hat{S}(x) = 1 - \hat{F}(x)$ .

**Example 5.3.3. Actuarial Exam Question.** The following is a sample of 10 payments:

4 4 5+ 5+ 5+ 8 10+ 10+ 12 15

where + indicates that a loss has exceeded the policy limit.

Using the Kaplan-Meier product-limit estimator, calculate the probability that the loss on a policy exceeds 11,  $\hat{S}(11)$ .

**Example Solution.** There are four event times (non-censored observations). For each time  $t_j$ , we can calculate the number of events  $s_j$  and the risk set  $R_j$  as the following:

$j$	$t_j$	$s_j$	$R_j$
1	4	2	10
2	8	1	5
3	12	1	2
4	15	1	1

Thus, the Kaplan-Meier estimate of  $S(11)$  is

$$\begin{aligned} \hat{S}(11) &= \prod_{j:t_j \leq 11} \left(1 - \frac{s_j}{R_j}\right) = \prod_{j=1}^2 \left(1 - \frac{s_j}{R_j}\right) \\ &= \left(1 - \frac{2}{10}\right) \left(1 - \frac{1}{5}\right) = (0.8)(0.8) = 0.64. \end{aligned}$$

**Example. 5.3.4. Bodily Injury Claims.** We consider again the Boston auto bodily injury claims data from Derrig et al. (2001) that was introduced in Example 5.1.11. In that example, we omitted the 17 claims that were censored by policy limits. Now, we include the full dataset and use the Kaplan-Meier product limit to estimate the survival function. This is given in Figure 5.5.

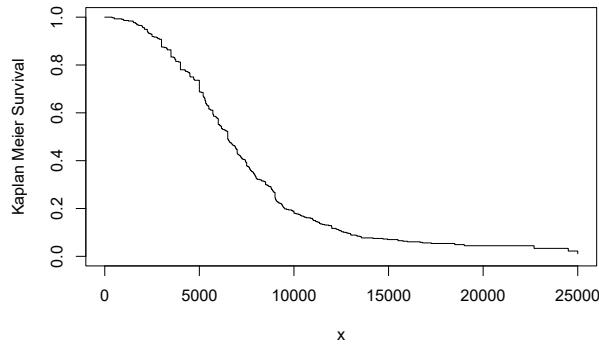


FIGURE 5.5: Kaplan-Meier Estimate of the Survival Function for Bodily Injury Claims

**Right-Censored, Left-Truncated Empirical Distribution Function.** In addition to right-censoring, we now extend the framework to allow for left-truncated data. As before, for each observation  $i$ , let  $u_i$  be the upper censoring limit ( $= \infty$  if no censoring). Further, let  $d_i$  be the lower truncation limit (0 if no truncation). Thus, the recorded value (if it is greater than  $d_i$ ) is  $x_i$  in the case of no censoring and  $u_i$  if there is censoring. Let  $t_1 < \dots < t_k$  be  $k$  distinct points at which an event of interest occurs, and let  $s_j$  be the number of recorded events  $x_i$ 's at time point  $t_j$ . The corresponding risk set is

$$R_j = \sum_{i=1}^n I(x_i \geq t_j) + \sum_{i=1}^n I(u_i \geq t_j) - \sum_{i=1}^n I(d_i \geq t_j).$$

With this new definition of the risk set, the product-limit estimator of the distribution function is as in equation (5.2).

**Greenwood's Formula.** (Greenwood, 1926) derived the formula for the estimated variance of the product-limit estimator to be

$$\widehat{Var}(\hat{F}(x)) = (1 - \hat{F}(x))^2 \sum_{j: t_j \leq x} \frac{s_j}{R_j(R_j - s_j)}.$$

As usual, we refer to the square root of the estimated variance as a *standard error*, a quantity that is routinely used in confidence intervals and for hypothesis testing. To compute this, R's `survfit` method takes a survival data object and creates a new object containing the Kaplan-Meier estimate of the survival function along with confidence intervals. The Kaplan-Meier method

(`type='kaplan-meier'`) is used by default to construct an estimate of the survival curve. The resulting discrete survival function has point masses at the observed event times (discharge dates)  $t_j$ , where the probability of an event given survival to that duration is estimated as the number of observed events at the duration  $s_j$  divided by the number of subjects exposed or 'at-risk' just prior to the event duration  $R_j$ .

**Alternative Estimators.** Two alternate types of estimation are also available for the `survfit` method. The alternative (`type='fh2'`) handles ties, in essence, by assuming that multiple events at the same duration occur in some arbitrary order. Another alternative (`type='fleming-harrington'`) uses the Nelson-Aalen (see (Aalen, 1978)) estimate of the **cumulative hazard function** to obtain an estimate of the survival function. The estimated cumulative hazard  $\hat{H}(x)$  starts at zero and is incremented at each observed event duration  $t_j$  by the number of events  $s_j$  divided by the number at risk  $R_j$ . With the same notation as above, the **Nelson-Äalen** estimator of the distribution function is

$$\hat{F}_{NA}(x) = \begin{cases} 0 & x < t_1 \\ 1 - \exp\left(-\sum_{j:t_j \leq x} \frac{s_j}{R_j}\right) & x \geq t_1 \end{cases}.$$

Note that the above expression is a result of the Nelson-Äalen estimator of the cumulative hazard function

$$\hat{H}(x) = \sum_{j:t_j \leq x} \frac{s_j}{R_j}$$

and the relationship between the survival function and cumulative hazard function,  $\hat{S}_{NA}(x) = e^{-\hat{H}(x)}$ .

**Example 5.3.5. Actuarial Exam Question.** For observation  $i$  of a survival study:

- $d_i$  is the left truncation point
- $x_i$  is the observed value if not right censored
- $u_i$  is the observed value if right censored

You are given:

Observation ( $i$ )	1	2	3	4	5	6	7	8	9	10
$d_i$	0	0	0	0	0	0	0	1.3	1.5	1.6
$x_i$	0.9	—	1.5	—	—	1.7	—	2.1	2.1	—
$u_i$	—	1.2	—	1.5	1.6	—	1.7	—	—	2.3

Calculate the Kaplan-Meier product-limit estimate,  $\hat{S}(1.6)$

**Example Solution.** Recall the risk set  $R_j = \sum_{i=1}^n \{I(x_i \geq t_j) + I(u_i \geq t_j) - I(d_i \geq t_j)\}$ . Then

$j$	$t_j$	$s_j$	$R_j$	$\hat{S}(t_j)$
1	0.9	1	$10 - 3 = 7$	$1 - \frac{1}{7} = \frac{6}{7}$
2	1.5	1	$8 - 2 = 6$	$\frac{6}{7} \left(1 - \frac{1}{6}\right) = \frac{5}{7}$
3	1.7	1	$5 - 0 = 5$	$\frac{5}{7} \left(1 - \frac{1}{5}\right) = \frac{4}{7}$
4	2.1	2	3	$\frac{4}{7} \left(1 - \frac{2}{3}\right) = \frac{4}{21}$

The Kaplan-Meier estimate is therefore  $\hat{S}(1.6) = \frac{5}{7}$ .

#### Example 5.3.6. Actuarial Exam Question. - Continued.

- a) Using the Nelson-Äalen estimator, calculate the probability that the loss on a policy exceeds 11,  $\hat{S}_{NA}(11)$ .
- b) Calculate Greenwood's approximation to the variance of the product-limit estimate  $\hat{S}(11)$ .

**Example Solution.** As before, there are four event times (non-censored observations). For each time  $t_j$ , we can calculate the number of events  $s_j$  and the risk set  $R_j$  as the following:

$j$	$t_j$	$s_j$	$R_j$
1	4	2	10
2	8	1	5
3	12	1	2
4	15	1	1

The Nelson-Aalen estimate of  $S(11)$  is  $\hat{S}_{NA}(11) = e^{-\hat{H}(11)} = e^{-0.4} = 0.67$ , since

$$\begin{aligned}\hat{H}(11) &= \sum_{j:t_j \leq 11} \frac{s_j}{R_j} = \sum_{j=1}^2 \frac{s_j}{R_j} \\ &= \frac{2}{10} + \frac{1}{5} = 0.2 + 0.2 = 0.4.\end{aligned}$$

From earlier work, the Kaplan-Meier estimate of  $S(11)$  is  $\hat{S}(11) = 0.64$ . Then Greenwood's estimate of the variance of the product-limit estimate of  $S(11)$  is

$$\widehat{Var}(\hat{S}(11)) = (\hat{S}(11))^2 \sum_{j:t_j \leq 11} \frac{s_j}{R_j(R_j - s_j)} = (0.64)^2 \left( \frac{2}{10(8)} + \frac{1}{5(4)} \right) = 0.0307.$$

## 5.4 Further Resources and Contributors

### Exercises

### Contributors

- **Zeinab Amin**, The American University in Cairo, is the principal author of this chapter. Email: [zeinabha@aucegypt.edu](mailto:zeinabha@aucegypt.edu) for chapter comments and suggested improvements.
- **Edward W. (Jed) Frees** and **Lisa Gao**, University of Wisconsin-Madison, are the principal authors of the sections on estimation using modified data which appeared in chapter 4 of the first edition of the text.
- Chapter reviewers include: Vytaras Brazauskas, Yvonne Chueh, Eren Dodd, Hirokazu (Iwahiro) Iwasawa, Joseph Kim, Andrew Kwon-Nakamura, Jian-dong Ren, and Di (Cindy) Xu.

### Further Readings and References

If you would like additional practice with R coding, please visit our companion [LDA Short Course](#). In particular, see the [Model Selection and Estimation Chapter](#).

# 6

---

## Model Selection

---

*Chapter Preview.* Model selection is a fundamental aspect of statistical modeling. In this chapter, the process of model selection is summarized, including tools for model comparisons and diagnostics. In addition to nonparametric tools for model selection based on marginal distributions of outcomes ignoring explanatory variables, this chapter underscores the idea that model selection is an iterative process in which models are cyclically (re)formulated and tested for appropriateness before using them for inference. After an overview, we describe the model selection process based on:

- an in-sample or training dataset,
- an out-of-sample or test dataset, and
- a method that combines these approaches known as cross-validation.

Although our focus is predominantly on data from continuous distributions, the same process can be used for discrete versions or data that come from a hybrid combination of discrete and continuous distributions.

---

In this chapter, you learn how to:

- Determine measures that summarize deviations of a parametric from a non-parametric fit
  - Describe the iterative model selection specification process
  - Outline steps needed to select a parametric model
  - Describe pitfalls of model selection based purely on in-sample data when compared to the advantages of out-of-sample model validation
- 

---

### 6.1 Tools for Model Selection and Diagnostics

Section 4.4.1 introduced nonparametric estimators in which there was no parametric form assumed about the underlying distributions. However, in many actuarial applications, analysts seek to employ a parametric fit of a distri-

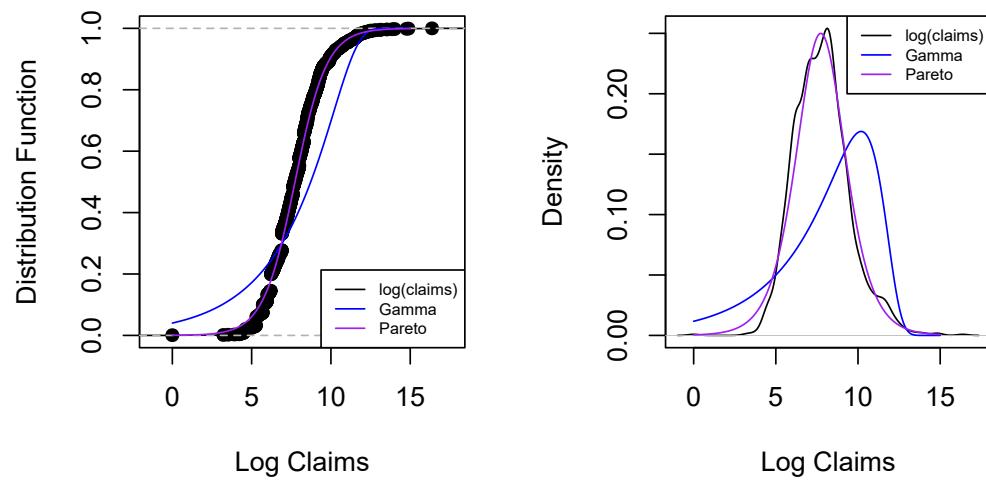
bution for ease of explanation and the ability to readily extend it to more complex situations such as including explanatory variables in a regression setting. When fitting a parametric distribution, one analyst might try to use a gamma distribution to represent a set of loss data. However, another analyst may prefer to use a Pareto distribution. How does one determine which model to select?

Nonparametric tools can be used to corroborate the selection of parametric models. Essentially, the approach is to compute selected summary measures under a fitted parametric model and to compare it to the corresponding quantity under the nonparametric model. As the nonparametric model does not assume a specific distribution and is merely a function of the data, it is used as a benchmark to assess how well the parametric distribution/model represents the data. Also, as the sample size increases, the empirical distribution converges almost surely to the underlying population distribution (by the strong law of large numbers). Thus the empirical distribution is a good proxy for the population. The comparison of parametric to nonparametric estimators may alert the analyst to deficiencies in the parametric model and sometimes point ways to improving the parametric specification. Procedures geared towards assessing the validity of a model are known as model diagnostics.

### 6.1.1 Graphical Comparison of Distributions

We have already seen the technique of overlaying graphs for comparison purposes. To reinforce the application of this technique, Figure 6.1 compares the empirical distribution to two parametric fitted distributions for log claims from the Property Fund data introduced in Section 1.3. The left panel shows the distribution functions of claims distributions. The dots forming an “S-shaped” curve represent the empirical distribution function at each observation. The thick blue curve gives corresponding values for the fitted gamma distribution and the light purple is for the fitted Pareto distribution. Because the Pareto is much closer to the empirical distribution function than the gamma, this provides evidence that the Pareto is the better model for this dataset. The right panel gives similar information for the density function and provides a consistent message. Based (only) on these figures, the Pareto distribution is the clear choice for the analyst.

For another way to compare the appropriateness of two fitted models, consider the probability-probability (*pp*) plot. A *pp* plot compares cumulative probabilities under two models. For our purposes, these two models are the nonparametric empirical distribution function and the parametric fitted model. Figure 6.2 shows *pp* plots for the Property Fund data. The fitted gamma is on the left and the fitted Pareto is on the right, compared to the same empirical



**FIGURE 6.1: Nonparametric Versus Fitted Parametric Distribution and Density Functions.** The left-hand panel compares distribution functions, with the dots corresponding to the empirical distribution, the thick blue curve corresponding to the fitted gamma and the light purple curve corresponding to the fitted Pareto. The right hand panel compares these three distributions summarized using probability density functions.

distribution function of the data. The straight line represents equality between the two distributions being compared, so points close to the line are desirable. As seen in earlier demonstrations, the Pareto is much closer to the empirical distribution than the gamma, providing additional evidence that the Pareto is the better model.

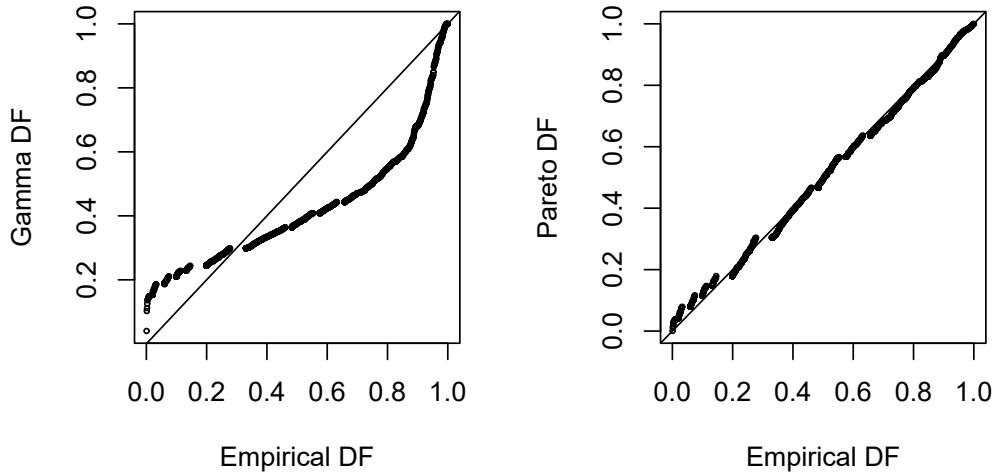
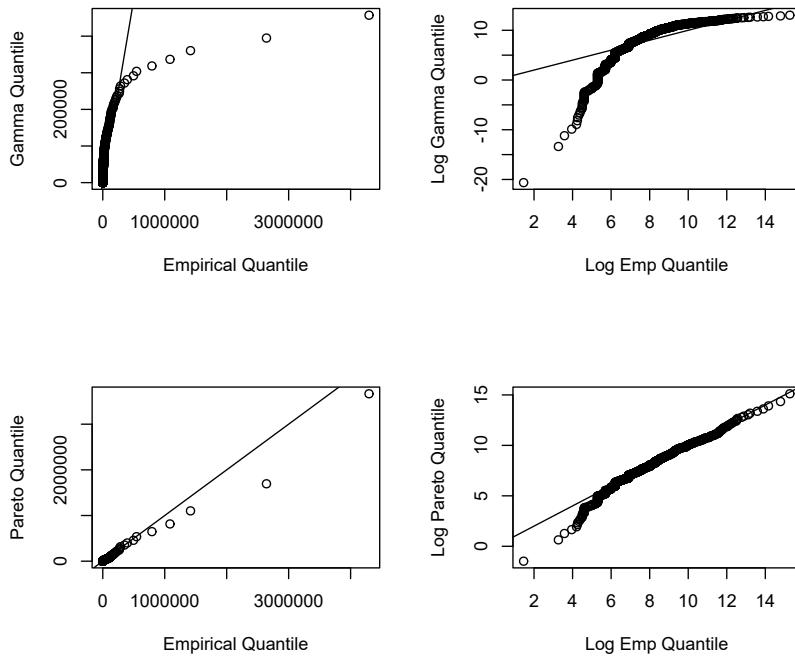


FIGURE 6.2: **Probability-Probability (*pp*) Plots.** The horizontal axis gives the empirical distribution function at each observation. In the left-hand panel, the corresponding distribution function for the gamma is shown in the vertical axis. The right-hand panel shows the fitted Pareto distribution. Lines of  $y = x$  are superimposed.

A *pp* plot is useful in part because no artificial scaling is required, such as with the overlaying of densities in Figure 6.1, in which we switched to the log scale to better visualize the data. Note further that *pp* plots are available in multivariate settings where more than one outcome variable is available. However, a limitation of the *pp* plot is that, because it plots *cumulative* distribution functions, it can sometimes be difficult to detect *where* a fitted parametric distribution is deficient. As an alternative, it is common to use a quantile-quantile (qq) plot, as demonstrated in Figure 6.3.

A *qq* plot compares two fitted models through their quantiles. As with *pp* plots, we compare the nonparametric to a parametric fitted model. Quantiles may be evaluated at each point of the dataset, or on a grid (e.g., at 0, 0.001, 0.002, ..., 0.999, 1.000), depending on the application. In Figure 6.3, for each point on the aforementioned grid, the horizontal axis displays the

empirical quantile and the vertical axis displays the corresponding fitted parametric quantile (gamma for the upper two panels, Pareto for the lower two). Quantiles are plotted on the original scale in the left panels and on the log scale in the right panels to allow us to see where a fitted distribution is deficient. The straight line represents equality between the empirical distribution and fitted distribution. From these plots, we again see that the Pareto is an overall better fit than the gamma. Furthermore, the lower-right panel suggests that the Pareto distribution does a good job with large claims, but provides a poorer fit for small claims.



**FIGURE 6.3: Quantile-Quantile ( $qq$ ) Plots.** The horizontal axis gives the empirical quantiles at each observation. The right-hand panels they are graphed on a logarithmic basis. The vertical axis gives the quantiles from the fitted distributions; gamma quantiles are in the upper panels, Pareto quantiles are in the lower panels.

**Example 6.1.1. Actuarial Exam Question.** Figure 6.4 shows a  $pp$  plot of a fitted distribution compared to a sample.

Comment on the two distributions with respect to left tail, right tail, and median probabilities.

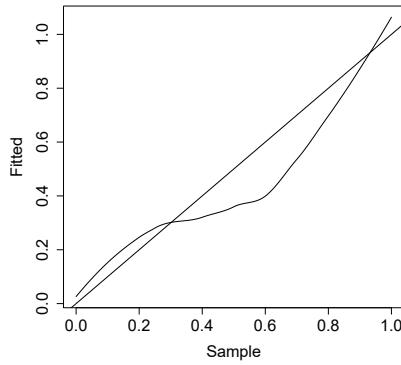


FIGURE 6.4: Example 6.1.1 Plot

**Example Solution.** The tail of the fitted distribution is too thick on the left, too thin on the right, and the fitted distribution has less probability around the median than the sample. To see this, recall that the *pp* plot graphs the cumulative distribution of two distributions on its axes (empirical on the x-axis and fitted on the y-axis in this case). For small values of  $x$ , the fitted model assigns greater probability to being below that value than occurred in the sample (i.e.  $F(x) > F_n(x)$ ). This indicates that the model has a heavier left tail than the data. For large values of  $x$ , the model again assigns greater probability to being below that value and thus less probability to being above that value (i.e.  $S(x) < S_n(x)$ ). This indicates that the model has a lighter right tail than the data. In addition, as we go from 0.4 to 0.6 on the horizontal axis (thus looking at the middle 20 data), the *pp* plot increases from about 0.3 to 0.4. This indicates that the model puts only about 10

### 6.1.2 Statistical Comparison of Distributions

When selecting a model, it is helpful to make the graphical displays presented. However, for reporting results, it can be effective to supplement the graphical displays with selected statistics that summarize model goodness of fit. Table 6.1 provides three commonly used goodness of fit statistics. In this table,  $F_n$  is the empirical distribution,  $F$  is the fitted or hypothesized distribution, and  $F_i^* = F(x_i)$ .

Table 6.1. Three Goodness of Fit Statistics

Statistic	Definition	Computational Expression
Kolmogorov-Smirnov	$\max_x  F_n(x) - F(x) $	$\max(D^+, D^-)$ where $D^+ = \max_{i=1,\dots,n} \left  \frac{i}{n} - F_i^* \right $ $D^- = \max_{i=1,\dots,n} \left  F_i^* - \frac{i-1}{n} \right $
Cramer-von Mises	$n \int (F_n(x) - F(x))^2 f(x) dx$	$\frac{1}{12n} + \sum_{i=1}^n (F_i^* - (2i-1)/n)^2$
Anderson-Darling	$n \int \frac{(F_n(x) - F(x))^2}{F(x)(1-F(x))} f(x) dx$	$-n - \frac{1}{n} \sum_{i=1}^n (2i-1) \log(F_i^*(1-F_{n+1-i}))^2$

The *Kolmogorov-Smirnov statistic* is the maximum absolute difference between the fitted distribution function and the empirical distribution function. Instead of comparing differences between single points, the *Cramer-von Mises statistic* integrates the difference between the empirical and fitted distribution functions over the entire range of values. The *Anderson-Darling statistic* also integrates this difference over the range of values, although weighted by the inverse of the variance. It therefore places greater emphasis on the tails of the distribution (i.e when  $F(x)$  or  $1 - F(x) = S(x)$  is small).

**Example 6.1.2. Actuarial Exam Question (modified).** A sample of claim payments is:

29 64 90 135 182

Compare the empirical claims distribution to an exponential distribution with mean 100 by calculating the value of the Kolmogorov-Smirnov test statistic.

r SolnBegin()‘ For an exponential distribution with mean 100, the cumulative distribution function is  $F(x) = 1 - e^{-x/100}$ . Thus,

x	$F(x)$	$F_n(x)$	$F_n(x-)$	$\max( F(x) - F_n(x) ,  F(x) - F_n(x-) )$
29	0.2517	0.2	0	$\max(0.0517, 0.2517) = 0.2517$
64	0.4727	0.4	0.2	$\max(0.0727, 0.2727) = 0.2727$
90	0.5934	0.6	0.4	$\max(0.0066, 0.1934) = 0.1934$
135	0.7408	0.8	0.6	$\max(0.0592, 0.1408) = 0.1408$
182	0.8380	1	0.8	$\max(0.1620, 0.0380) = 0.1620$

The Kolmogorov-Smirnov test statistic is therefore

$$KS = \max(0.2517, 0.2727, 0.1934, 0.1408, 0.1620) = 0.2727.$$

r SolnEnd()‘

### Pearson's chi-square test

In this section we introduce another goodness of fit test - Pearson's chi-square test - which can be used for testing whether a discrete distribution provides a good fit to discrete data. For more details on the Pearson's chi-square test, at an introductory mathematical statistics level, we refer the reader to Section 9.1 of [Hogg et al. \(2015\)](#).

To illustrate application of the Pearson's chi-square test, we use the example introduced in Section 3.7: In 1993, a portfolio of  $n = 7,483$  automobile insurance policies from a major Singaporean insurance company had the distribution of auto accidents per policyholder as given in [Table 6.2](#).

**Table 6.2. Singaporean Automobile Accident Data**

Count ( $k$ )	0	1	2	3	4	Total
No. of Policies with $k$ accidents ( $m_k$ )	6,996	455	28	4	0	7,483

If we fit a Poisson distribution, then the *mle* for  $\lambda$ , the Poisson mean, is the sample mean which is given by

$$\bar{N} = \frac{0 \cdot 6996 + 1 \cdot 455 + 2 \cdot 28 + 3 \cdot 4 + 4 \cdot 0}{7483} = 0.06989.$$

Now if we use Poisson ( $\hat{\lambda}_{MLE}$ ) as the fitted distribution, then a tabular comparison of the fitted counts and observed counts is given by [Table 6.3](#), where  $\hat{p}_k$  represents the estimated probabilities under the fitted Poisson distribution.

**Table 6.3. Comparison of Observed to Fitted Counts: Singaporean Auto Data**

Count ( $k$ )	Observed ( $m_k$ )	Fitted Counts Using Poisson ( $n\hat{p}_k$ )
0	6,996	6,977.86
1	455	487.70
2	28	17.04
3	4	0.40
$\geq 4$	0	0.01
Total	7,483	7,483.00

While the fit seems reasonable, the Pearson's chi-square statistic is a goodness of fit measure that can be used to test the hypothesis that the underlying distribution is Poisson. To explain this statistic let us suppose that a dataset of size  $n$  is grouped into  $k$  cells with  $m_k/n$  and  $\hat{p}_k$ , for  $k = 1 \dots, K$  being the

observed and estimated probabilities of an observation belonging to the  $k$ -th cell, respectively. The Pearson's chi-square test statistic is then given by

$$\sum_{k=1}^K \frac{(m_k - n\hat{p}_k)^2}{n\hat{p}_k}.$$

The motivation for the above statistic derives from the fact that

$$\sum_{k=1}^K \frac{(m_k - np_k)^2}{np_k}$$

has a limiting chi-square distribution with  $K - 1$  degrees of freedom if  $p_k$ ,  $k = 1, \dots, K$  are the true cell probabilities. Now suppose that only the summarized data represented by  $m_k$ ,  $k = 1, \dots, K$  is available. Further, if  $p_k$ 's are functions of  $s$  parameters, replacing  $p_k$ 's by any *efficiently* estimated probabilities  $\hat{p}_k$ 's results in the statistic continuing to have a limiting chi-square distribution but with degrees of freedom given by  $K - 1 - s$ . Such efficient estimates can be derived for example by using the *mle* method (with a multinomial likelihood) or by estimating the  $s$  parameters which minimizes the Pearson's chi-square statistic above. For example, the R code below does calculate an estimate for  $\lambda$  doing the latter and results in the estimate 0.06623153, close but different from the *mle* of  $\lambda$  using the full data:

```
m <- c(6996, 455, 28, 4, 0);
op <- m/sum(m);
g <- function(lam){ sum( (op-c(dpois(0:3, lam), 1-ppois(3, lam))) )^2 };
optim( sum(op*(0:4)), g, method="Brent", lower=0, upper=10)$par
```

When one uses the full data to estimate the probabilities, the asymptotic distribution is *in between* chi-square distributions with parameters  $K - 1$  and  $K - 1 - s$ . In practice it is common to ignore this subtlety and assume the limiting chi-square has  $K - 1 - s$  degrees of freedom. Interestingly, this practical shortcut works quite well in the case of the Poisson distribution.

For the Singaporean auto data the Pearson's chi-square statistic equals 41.98 using the full data *mle* for  $\lambda$ . Using the limiting distribution of chi-square with  $5 - 1 - 1 = 3$  degrees of freedom, we see that the value of 41.98 is way out in the tail (99-th percentile is below 12). Hence we can conclude that the Poisson distribution provides an inadequate fit for the data.

In the above, we started with the cells as given in the above tabular summary. In practice, a relevant question is how to define the cells so that the chi-square distribution is a good approximation to the finite sample distribution of the statistic. A rule of thumb is to define the cells in such a way to have at least 80%, if not all, of the cells having expected counts greater than 5. Also, it is

clear that a larger number of cells results in a higher power of the test, and hence a simple rule of thumb is to maximize the number of cells such that each cell has at least 5 observations.

---

## 6.2 Iterative Model Selection

In our model development, we examine the data graphically, hypothesize a model structure, and compare the data to a candidate model in order to formulate an improved model. Box (1980) describes this as an *iterative process* which is shown in Figure 6.5.

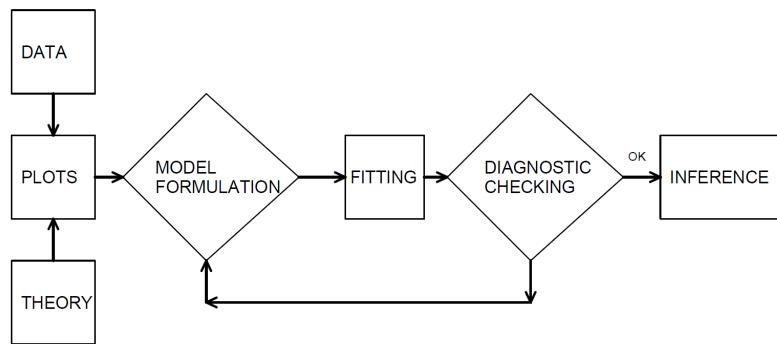


FIGURE 6.5: Iterative Model Specification Process

This iterative process provides a useful recipe for structuring the task of specifying a model to represent a set of data.

1. The first step, the model formulation stage, is accomplished by examining the data graphically and using prior knowledge of relationships, such as from economic theory or industry practice.
2. The second step in the iteration is fitting based on the assumptions of the specified model. These assumptions must be consistent with the data to make valid use of the model.
3. The third step is *diagnostic checking*; the data and model must be consistent with one another before additional inferences can be made. Diagnostic checking is an important part of the model formulation; it can reveal mistakes made in previous steps and provide ways to correct these mistakes.

The iterative process also emphasizes the skills you need to make data analytics work. First, you need a willingness to summarize information numerically

and portray this information graphically. Second, it is important to develop an understanding of model properties. You should understand how a probabilistic model behaves in order to match a set of data to it. Third, theoretical properties of the model are also important for inferring general relationships based on the behavior of the data.

---

### 6.3 Model Selection Based on a Training Dataset

As introduced in Section 2.2, it is common to refer to a dataset used for fitting the model as a *training* or an *in-sample* dataset. Techniques available for selecting a model depend upon whether the outcomes  $X$  are discrete, continuous, or a hybrid of the two, although the principles are the same.

**Graphical and other Basic Summary Measures.** Begin by summarizing the data graphically and with statistics that do not rely on a specific parametric form, as summarized in Section 4.4.1. Specifically, you will want to graph both the empirical distribution and density functions. Particularly for loss data that contain many zeros and that can be skewed, deciding on the appropriate scale (e.g., logarithmic) may present some difficulties. For discrete data, tables are often preferred. Determine sample moments, such as the mean and variance, as well as selected quantiles, including the minimum, maximum, and the median. For discrete data, the mode (or most frequently occurring value) is usually helpful.

These summaries, as well as your familiarity of industry practice, will suggest one or more candidate parametric models. Generally, start with the simpler parametric models (for example, one parameter exponential before a two parameter gamma), gradually introducing more complexity into the modeling process.

Critique the candidate parametric model numerically and graphically. For the graphs, utilize the tools introduced in Section 6.1 such as  $pp$  and  $qq$  plots. For the numerical assessments, examine the statistical significance of parameters and try to eliminate parameters that do not provide additional information. In addition to statistical significance of parameters, you may use the following model comparison tools.

**Likelihood Ratio Tests.** For comparing model fits, if one model is a subset of another, then a likelihood ratio test may be employed; the general approach to likelihood ratio testing is described in Appendix Sections ?? and ??.

**Goodness of Fit Statistics.** Generally, models are not proper subsets of

one another in which case overall goodness of fit statistics are helpful for comparing models. *Information criteria* are one type of goodness of statistic. The most widely used examples are Akaike's Information Criterion (*AIC*) and the (Schwarz) Bayesian Information Criterion (*BIC*); they are widely cited because they can be readily generalized to multivariate settings. Appendix Section ?? provides a summary of these statistics.

For selecting the appropriate distribution, statistics that compare a parametric fit to a nonparametric alternative, summarized in Section 6.1.2, are useful for model comparison. For discrete data, a *goodness of fit* statistic is generally preferred as it is more intuitive and simpler to explain.

---

## 6.4 Model Selection Based on a Test Dataset

Model validation introduced in Section 2.2 is the process of confirming that the proposed model is appropriate based on a *test* or an *out-of-sample* dataset, especially in light of the purposes of the investigation. Model validation is important since the model selection process based only on training or in-sample data can be susceptible to data-snooping, that is, fitting a great number of models to a single set of data. By looking at a large number of models, we may overfit the data and underestimate the natural variation in our representation.

Selecting a model based only on in-sample data also does not support the goal of predictive inference. Particularly in actuarial applications, our goal is to make statements about *new* experience rather than a dataset at hand. For example, we use claims experience from one year to develop a model that can be used to price insurance contracts for the following year. As an analogy, we can think about the training dataset as experience from one year that is used to predict the behavior of the next year's test dataset.

We can respond to these criticisms by using a technique known as *out-of-sample validation*. The ideal situation is to have available two sets of data, one for training, or model development, and the other for testing, or model validation. We initially develop one or several models on the first dataset that we call *candidate* models. Then, the relative performance of the candidate models can be measured on the second set of data. In this way, the data used to validate the model are unaffected by the procedures used to formulate the model.

**Random Split of the Data.** Unfortunately, rarely will two sets of data be available to the investigator. As mentioned in Section 2.2, we can implement

the validation process by splitting the dataset into *training* and *test* subsamples, respectively. Figure 6.6 illustrates this splitting of the data.

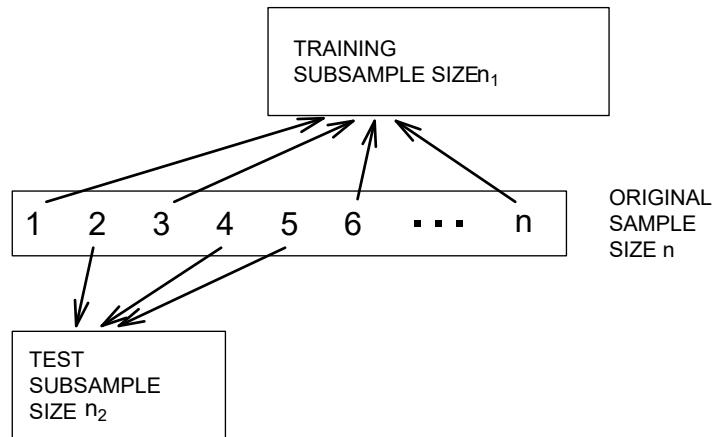


FIGURE 6.6: **Model Validation.** A dataset is randomly split into two subsamples.

Various researchers recommend different proportions for the allocation. [Snee \(1977\)](#) suggests that data-splitting not be done unless the sample size is moderately large. The guidelines of [Picard and Berk \(1990\)](#) show that the greater the number of parameters to be estimated, the greater the proportion of observations is needed for the training subsample for model development.

**Selecting a Distribution.** Still, our focus so far has been to select a distribution for a dataset that can be used for actuarial modeling without additional explanatory or input variables  $x_1, \dots, x_k$ . Even in this more fundamental problem, the model validation approach is valuable. If we base all inference on only in-sample data, then there is a tendency to select more complicated models than needed. For example, we might select a four parameter GB2, generalized beta of the second kind, distribution when only a two parameter Pareto is needed. Information criteria such as AIC and BIC introduced in Appendix Section ?? include penalties for model complexity and thus provide protection against over-fitting, but using a test sample may also help achieve parsimonious models. From a quote often attributed to Albert Einstein, we want to “use the simplest model as possible but no simpler.”

---

**Example 6.4.1. Wisconsin Property Fund.** For the 2010 property fund

data from Section 1.3, we may try to select a severity distribution based on out-of-sample prediction. In particular, we may randomly select 1,000 observations as our training data, and use the remaining 377 claims to validate the two models based respectively on gamma and Pareto distributions. For illustration purposes, We compare the Kolmogorov-Smirnov statistics respectively for the training and test datasets using the models fitted from training data.

Based on in-sample prediction, the Kolmogorov-Smirnov goodness of fit statistic for the gamma distribution turns out to be 0.2771 and for the Pareto distribution is 0.046. Based on out-of-sample prediction, the Kolmogorov-Smirnov goodness of fit statistic for the gamma distribution turns out to be 0.2693 and for the Pareto distribution is 0.0746. Based on both in-sample and out-of-sample prediction, the Pareto model seems to give considerably better goodness of fit under the random seed used in the code for splitting the training and test data.

**Model Validation Statistics.** In addition to the nonparametric tools introduced earlier for comparing marginal distributions of the outcome or output variables ignoring potential explanatory or input variables, much of the literature supporting the establishment of a model validation process is based on regression and classification models that you can think of as an *input-output* problem (James et al. (2013)). That is, we have several inputs or predictor variables  $x_1, \dots, x_k$  that are related to an output or outcome  $y$  through a function such as

$$y = g(x_1, \dots, x_k).$$

For model selection, one uses the training sample to develop an estimate of  $g$ , say,  $\hat{g}$ , and then calibrate the average distance from the observed outcomes to the predictions using a criterion of the form

$$\frac{1}{n} \sum_i d(y_i, \hat{g}(x_{i1}, \dots, x_{ik})). \quad (6.1)$$

Here, “d” is some measure of distance and the sum  $i$  is over the test data. The function  $g$  may not have an analytical form and can be estimated for each observation using the different different types of algorithms and models introduced earlier in Section 2.4. In many regression applications, it is common to use the squared Euclidean distance of the form  $d(y_i, g) = (y_i - g)^2$  under which the criterion in equation (6.1) is called the *mean squared error (MSE)*. Using data simulated from linear models, Example 2.3.1 uses the *root mean squared error (Rmse)* which is the squared root of the MSE. From equation (6.1), the MSE criteria works the best for linear models under normal distributions with constant variance, as minimizing MSE is equivalent to the maximum likelihood and least squares criterion in training data. In data analytics and linear

regression, one may consider transformations of the outcome variable in order for the MSE criteria to work more effectively. In actuarial applications, the *mean absolute error (MAE)* under the Euclidean distance  $d(y_i, g) = |y_i - g|$  may be preferred because of the skewed nature of loss data. For right-skewed outcomes, it may require a larger sample size for the validation statistics to pickup the correct model when large outlying values of  $y$  can have a large effect on the measures.

Following Example 2.3.1, we use simulated data in Examples 6.4.2 through 6.4.4 to compare the AIC information criteria from Appendix Chapter ?? with out-of-sample MSE and MAE criterion for selecting the distribution and input variables for outcomes that are respectively from normal and right-skewed distributions including lognormal and gamma distributions. For right skewed distributions, we find that the AIC information criteria seems to work consistently for selecting the correct distributional form and mean structure (input variables), whereas out-of-sample MSE and MAE may not work for right-skewed outcomes like those from gamma distributions, even with relatively large sample sizes. Therefore, model validation statistics commonly used in data analytics may only work for minimizing specific cost functions, such as the MAE that represents the average absolute error for out-of-sample prediction, and do not necessarily guarantee correct selection of the underlying data generating mechanism.

**Example 6.4.2. In-sample AIC and out-of-sample MSE for normal outcomes.** Example 2.3.1 assumes that there is a set of claims that potentially varies by a single categorical variable with six levels. To illustrating in-sample over-fitting, it also assumes that two of the six levels share a common mean that differs from rest of levels. For Example 2.3.1, the claim amounts were generated from a linear model with constant variance, for which in-sample AIC and out-of-sample Rmse provide consistent results from the cross-validation procedure to be introduced in the next section. Here, we may use the same data generation mechanism to compare the performance of in-sample AIC with the in-sample and out-of-sample Rmse criteria. In particular, we generate a total of 200 samples and split them equally into the training and test datasets. From Table 6.4, we observe the two-level model was correctly selected by both in-sample AIC and out-of-sample MSE criteria, whereas in-sample MSE prefers an over-fitted model with six levels. Thus, due to concerns of model overfitting, we do not use in-sample distance measures such as the MSE and MAE criterion that favors more complicated models.

TABLE 6.4: Model Selection based on MSE and AIC for normal outputs

	Community Rating	Two Levels	Six Levels
Rmse - Train	1.186	1.016	0.990
Rmse - Test	1.081	0.958	1.012
AIC - Train	321.935	293.028	295.694

**Example 6.4.3. MSE and MAE for right-skewed outcomes - lognormal claims.** For claims modeling, one may wonder how the MSE and MAE types of criterion may perform for right-skewed data. Using the same data generating procedure, we may generate lognormal claim amounts by exponentiating the normal outcomes from the previous example. We fit the lognormal claim amounts with lognormal and gamma regression commonly used for ratemaking and claims analytics. Results are summarized in Tables 6.5 and 6.6, respectively. For the specific data generating mechanism, we observe that it requires a larger sample size for out-of-sample Rmse and MAE to select the correct distributional form and mean structure, when compared with in-sample AIC criteria. The AIC criteria is able to pick out the correct model with a sample size of 200, while out-of-sample MSE and MAE fail to. Thus, for right skewed output, precautions need to be taken when using model validation statistics that may be sensitive to large claim values, particularly when the sample size is relatively small.

**TABLE 6.5: Model Selection based on in-sample AIC and out-of-sample MSE and MAE from lognormal model**

	Community Rating	Two Levels	Six Levels
Rmse - Train	4.365	4.185	4.192
Rmse - Test	3.881	3.686	3.679
MAE - Train	2.077	1.821	1.807
MAE - Test	2.166	2.056	2.073
AIC - Train	1800.716	1681.550	1686.142

**TABLE 6.6: Model Selection based on in-sample AIC and out-of-sample MSE and MAE from gamma model**

	Community Rating	Two Levels	Six Levels
Rmse - Train	4.634	4.572	4.572
Rmse - Test	4.298	4.232	4.235
MAE - Train	1.862	1.815	1.817
MAE - Test	2.127	2.123	2.128
AIC - Train	1906.398	1789.312	1795.662

---

**Example 6.4.4. MSE and MAE for right-skewed outcomes - gamma claims.** For right-skewed outcomes, we may be interested in studying how the MSE and MAE types of measures work for another loss severity distribution, the gamma distribution, that is widely used in ratemaking and claims analytics. Here, we use a similar mean structure for generating claims amounts based on a gamma regression with the log link function. We fit the data using lognormal and gamma regression. Results are summarized in Tables 6.7 and 6.8, respectively. For gamma outcomes, Table 6.8 shows that out-of-sample MSE and MAE criterion fail to select the correct distributional form or the mean structure even with a total of 1000 samples. By changing the gamma shape parameter, you may see that the out-of-sample MSE and MAE criterion work in certain settings for correctly selecting the distributional form or the mean structure, but the performance of such model validation statistics does not seem to be consistent across different parameter values and sample sizes for right-skewed gamma outcomes. Again, the AIC criteria seems to be working consistently in selecting the correct distribution and mean structure for the data generated from gamma distributions, even with a smaller sample size of 200.

TABLE 6.7: Model Selection based on in-sample AIC and out-of-sample MSE and MAE from lognormal model

	Community Rating	Two Levels	Six Levels
Rmse - Train	1.083	0.763	0.760
Rmse - Test	1.128	0.815	0.812
MAE - Train	0.800	0.535	0.529
MAE - Test	0.830	0.565	0.566
AIC - Train	1212.218	864.776	868.794

TABLE 6.8: Model Selection based on in-sample AIC and out-of-sample MSE and MAE from gamma model

	Community Rating	Two Levels	Six Levels
Rmse - Train	1.553	1.476	1.475
Rmse - Test	1.594	1.523	1.522
MAE - Train	1.121	1.226	1.227
MAE - Test	1.138	1.253	1.253
AIC - Train	1249.211	852.292	856.850

## 6.5 Model Selection Based on Cross-Validation

Although out-of-sample validation is the gold standard in predictive modeling, it is not always practical to do so. The main reason is that we have limited sample sizes and the out-of-sample model selection criterion in equation (6.1) depends on a *random* split of the data. This means that different analysts, even when working the same dataset and same approach to modeling, may select different models. This is likely in actuarial applications because we work with skewed datasets where there is a large chance of getting some very large outcomes and large outcomes may have a great influence on the parameter estimates.

**Cross-Validation Procedure.** Alternatively, one may use *cross-validation*, as follows.

- The procedure begins by using a random mechanism to split the data into  $K$  subsets of roughly equal size known as *folds*, where analysts typically use 5 to 10.
- Next, one uses the first  $K-1$  subsamples to estimate model parameters. Then,

“predict” the outcomes for the  $K$ th subsample and use a measure such as in equation (6.1) to summarize the fit.

- Now, repeat this by holding out each of the  $K$  subsamples, summarizing with an out-of-sample statistic. Thus, summarize these  $K$  statistics, typically by averaging, to give a single overall statistic for comparison purposes.

Repeat these steps for several candidate models and choose the model with the lowest overall cross-validation statistic.

In Example 2.3.1, you have seen that the MSE criteria seems to work with  $k$ -fold cross-validation in selecting the correct mean structure for claims outcome data generated from linear models with constant variance. From Examples 6.4.3 and 6.4.4, however, the out-of-sample MSE and MAE criterion does not seem to provide consistent performance for selecting the distributional form and the mean structure under right-skewed claims distributions. Thus, we may use the  $k$ -folder cross-validation instead of out-of-sample prediction to see whether the MSE and MAE types of criterion work for right-skewed distributions based on lognormal and gamma regression with a log link function.

**Example 6.5.1. Cross-validation in right-skewed outcomes - lognormal claims** For lognormal claims, we use the data generating mechanism from Example 6.4.3 to generate a total of 100 samples, and use the  $k$ -fold cross validation procedure in Example 2.3.1 to select the distributional form and mean structure. Using cross-validation, we note that both AIC and out-of-sample MSE and MAE seem to be working for selecting the model with the correct distribution and mean structure, even with a total of 100 samples.

**Example 6.5.2. Cross-validation in right-skewed outcomes - gamma claims** For gamma claims, we use the data generating mechanism from Example 6.4.4 to generate a total of 100 samples, and use the  $k$ -fold cross validation procedure to select the distributional form and mean structure. Using cross-validation, we note that in-sample AIC seems to be working for selecting the model with the correct distribution and mean structure, while out-of-sample MSE and MAE seem to fail in selecting the distributional form or the mean structure correctly even after we increase the sample size to 1000.

Cross-validation is widely used because it retains the predictive flavor of the out-of-sample model validation process but, due to the re-use of the data, is more stable over random samples. In addition, Example 8.4.1 in Chapter 8 uses the Wisconsin Property Fund to perform  $k$ -fold cross-validation of the gamma and Pareto models based on the Kolmogorov-Smirnov goodness of fit

TABLE 6.9: Cross-validation based on in-sample AIC, and out-of-sample MSE and MAE from lognormal model

	Community Rating	Two Levels	Six Levels
Rmse - Fold 1	1.808	1.750	1.891
Rmse - Fold 2	2.145	1.773	1.813
Rmse - Fold 3	3.461	3.335	3.333
Rmse - Fold 4	1.425	1.723	1.865
Rmse - Fold 5	4.848	4.450	4.454
Rmse - Average	2.738	2.606	2.671
MAE - Fold 1	1.341	1.408	1.502
MAE - Fold 2	1.881	1.264	1.255
MAE - Fold 3	2.037	2.142	2.146
MAE - Fold 4	1.225	1.345	1.476
MAE - Fold 5	2.421	2.022	2.051
MAE - Average	1.781	1.636	1.686
AIC - Average	286.257	266.223	271.200

TABLE 6.10: Cross-validation based on in-sample AIC, and out-of-sample MSE and MAE from gamma model

	Community Rating	Two Levels	Six Levels
Rmse - Fold 1	2.557	2.642	2.677
Rmse - Fold 2	1.930	1.999	2.005
Rmse - Fold 3	4.088	4.155	4.187
Rmse - Fold 4	1.181	1.273	1.318
Rmse - Fold 5	5.232	5.262	5.286
Rmse - Average	2.998	3.066	3.095
MAE - Fold 1	1.929	2.069	2.114
MAE - Fold 2	1.060	1.116	1.124
MAE - Fold 3	2.488	2.660	2.725
MAE - Fold 4	0.887	0.949	0.999
MAE - Fold 5	2.251	2.312	2.345
MAE - Average	1.723	1.821	1.861
AIC - Average	299.063	281.455	282.816

TABLE 6.11: Cross-validation based on in-sample AIC, and out-of-sample MSE and MAE from lognormal model

	Community Rating	Two Levels	Six Levels
Rmse - Fold 1	1.080	0.794	0.799
Rmse - Fold 2	0.953	0.639	0.639
Rmse - Fold 3	1.354	0.914	0.916
Rmse - Fold 4	1.097	0.725	0.727
Rmse - Fold 5	1.171	0.695	0.695
Rmse - Average	1.131	0.753	0.755
MAE - Fold 1	0.837	0.579	0.583
MAE - Fold 2	0.755	0.473	0.474
MAE - Fold 3	0.952	0.600	0.602
MAE - Fold 4	0.852	0.523	0.525
MAE - Fold 5	0.897	0.503	0.507
MAE - Average	0.859	0.536	0.538
AIC - Average	1980.018	1381.321	1388.351

TABLE 6.12: Cross-validation based on in-sample AIC, and out-of-sample MSE and MAE from gamma model

	Community Rating	Two Levels	Six Levels
Rmse - Fold 1	1.455	1.620	1.620
Rmse - Fold 2	1.347	1.543	1.543
Rmse - Fold 3	1.865	2.006	2.005
Rmse - Fold 4	1.558	1.738	1.738
Rmse - Fold 5	1.690	1.838	1.838
Rmse - Average	1.583	1.749	1.749
MAE - Fold 1	1.003	1.223	1.223
MAE - Fold 2	0.975	1.195	1.195
MAE - Fold 3	1.301	1.478	1.479
MAE - Fold 4	1.118	1.342	1.342
MAE - Fold 5	1.228	1.420	1.420
MAE - Average	1.125	1.332	1.332
AIC - Average	2047.108	1349.855	1357.246

statistic. Additional information and examples regarding re-sampling procedures including leave-one-out cross-validation and bootstrap can also be found in Chapter 8.

---

## 6.6 Model Selection for Modified Data

So far we have discussed model selection using unmodified data. For modified data including grouped, censored and truncated data, you learned parametric and nonparametric estimation of distribution functions in Section 5.2. For model selection, the tools from Section 6.1 can be extended to cases of modified data.

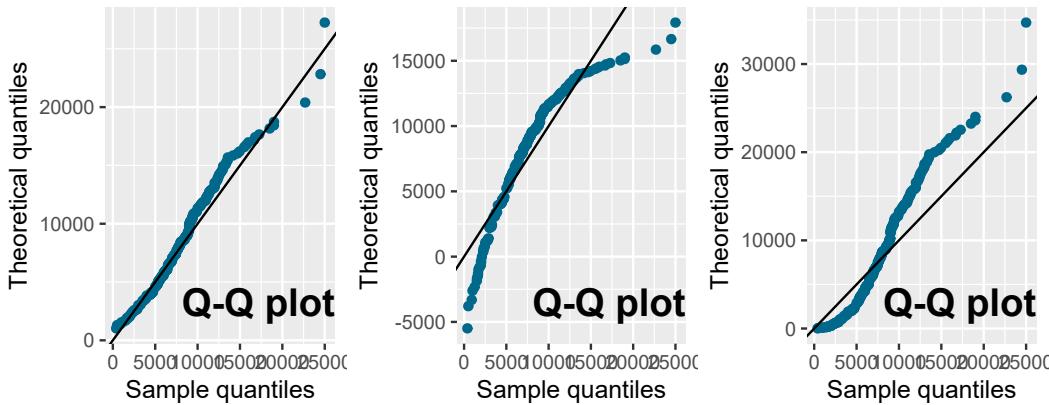
For selection of distributions, the nonparametric tools introduced in Section 6.1 are based on estimated parametric and nonparametric distribution functions, and thus can be extended to modified data for which both types of estimators exist.

For graphical comparisons, the *pp* and *qq* plots introduced earlier can be created for modified data by plotting the parametric estimates from Section 5.2 against nonparametric estimates of the probability or distribution functions from Section 5.3. For example, the `qqPlotCensored` and `qqtrunc` functions in R generate *qq* plots respectively for censored (left or right) and truncated data, whereas the `probPlot` function creates both *pp* and *qq* plots with a larger selection of distributions for right-censored and unmodified data. Additional graphical tools such as cumulative hazard plots are available in the R package `GofCens`.

---

**Example 6.6.1. Bodily Injury Claims and *qq*-Plots.** For the Boston auto bodily injury claims data from Example 5.3.2, we include the full dataset with right-censoring, and use the *qq*-plot to compare the estimated quantiles from lognormal, normal and exponential distributions with those from the nonparametric Kaplan-Meier method. From the *qq*-plots in Figure 6.7, the lognormal distribution seems to fit the censored data much better than those based on the normal and exponential distributions.

In addition to graphical tools, you may use tools from Section 6.1.2 for statistical comparisons of models fitted from modified data based on parametric and nonparametric estimates of distribution functions. For example, the R package `GofCens` provides functions calculating the three goodness of fit statistics from Section 6.1.2 for both right-censored and unmodified data. The R pack-



**FIGURE 6.7: Quantile-Quantile ( $qq$ ) Plots for Bodily Injury Claims.** The horizontal axis gives the empirical quantiles at each observation. The vertical axis gives the quantiles from the fitted distributions; lognormal quantiles are in the left-hand panel, normal quantiles are in the middle, and exponential in the right-hand panel.

**TABLE 6.13: Nonparametric goodness of fit statistics for right-censored Bodily Injury Claims**

	Kolmogorov-Smirnov	Cramer-von Mises	Anderson-Darling
Lognormal	1.994	0.305	1.770
Normal	3.096	1.335	9.437
Exponential	4.811	4.065	21.659

age `truncgof`, on the other hand, provides functions for calculating the three goodness of fit statistics for left-truncated data.

**Example 6.6.2. Bodily Injury Claims and Goodness of Fit Stastistics.** For the Boston auto bodily injury claims with right-censoring, we may use the goodness of fit statistics to evaluate the fitted lognormal, normal and exponential distributions. For the Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling statistics, the lognormal distribution gives values that are much lower than those from normal and exponential distributions. The conclusion from the goodness of fit statistics is consistent to that revealed by the  $qq$  plots.

Other than selecting the distributional form, model comparison measures such as the likelihood ratio test and information criterion including the AIC from Section 6.3 can be obtained for models fitted based on likelihood criteria based on the likelihood functions introduced earlier for modified data. For modified

data, the `survreg` and `flexsurvreg` functions in R fit parametric regression models on censored and/or truncated outcomes based on maximum likelihood estimation which allows use of likelihood ratio tests and information criterion such as AIC for in-sample model comparisons. For censored and truncated data, the functions also provide output of residuals that allow calculation of model validation statistics such as the MSE and MAE for the iterative model selection procedure introduced in Section 6.2.

---

## 6.7 Further Resources and Contributors

### Contributors

- **Lei (Larry) Hua** and **Michelle Xia**, Northern Illinois University, are the principal authors of the second edition of this chapter.
- **Edward (Jed) Frees** and **Lisa Gao**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter.
- Chapter reviewers include: Vytaras Brazauskas, Yvonne Chueh, Eren Dodd, Hirokazu (Iwahiro) Iwasawa, Joseph Kim, Andrew Kwon-Nakamura, Jian-dong Ren, and Di (Cindy) Xu.

### Further Readings and References

If you would like additional practice with R coding, please visit our companion [LDA Short Course](#). In particular, see the [Model Selection and Estimation Chapter](#).

# 7

---

## Aggregate Loss Models

---

*Chapter Preview.* This chapter introduces probability models for describing the aggregate (total) claims that arise from a portfolio of insurance contracts. We present two standard modeling approaches, the individual risk model and the collective risk model. Further, we discuss strategies for computing the distribution of the aggregate claims, including exact methods for special cases, recursion, and simulation. Finally, we examine the effects of individual policy modifications such as deductibles, coinsurance, and inflation, on the frequency and severity distributions, and thus on the aggregate loss distribution.

---

### 7.1 Introduction

---

In this section, you learn:

- the concept of aggregate claims for an insurance system
  - alternative methods to describe the aggregate losses
  - the interpretation of different models for aggregate claims
- 

The objective of this chapter is to build a probability model to describe the aggregate claims by an insurance system occurring in a fixed time period. The insurance system could be a single policy, a group insurance contract, a business line, or an entire book of an insurer's business. In this chapter, aggregate claims refer to either the number or the amount of claims from a portfolio of insurance contracts. However, the modeling framework can be readily applied in a general setup.

Consider an insurance portfolio of  $n$  individual contracts, and let  $S$  denote the aggregate losses of the portfolio in a given time period. There are two approaches to modeling the aggregate losses  $S$ , the individual risk model and the collective risk model. The individual risk model emphasizes the loss from

each individual contract and represents the aggregate losses as:

$$S_n = X_1 + X_2 + \cdots + X_n,$$

where  $X_i$  ( $i = 1, \dots, n$ ) is interpreted as the loss amount from the  $i$ th contract. It is worth stressing that  $n$  denotes the number of contracts in the portfolio and thus is a fixed number rather than a random variable. For the individual risk model, one usually assumes the  $X_i$ 's are independent. Because of different contract features such as coverage and exposure, the  $X_i$ 's are not necessarily identically distributed. A notable feature of the distribution of each  $X_i$  is the probability mass at zero corresponding to the event of no claims.

The collective risk model represents the aggregate losses in terms of a frequency distribution and a severity distribution:

$$S_N = X_1 + X_2 + \cdots + X_N.$$

Here, one thinks of a random number of claims  $N$  that may represent either the number of losses or the number of payments. In contrast, in the individual risk model, we use a fixed number of contracts  $n$ . We think of  $X_1, X_2, \dots, X_N$  as representing the amount of each loss. Each loss may or may not correspond to a unique contract. For instance, there may be multiple claims arising from a single contract. It is natural to think about  $X_i > 0$  because if  $X_i = 0$  then no claim has occurred. Typically we assume that conditional on  $N = n$ ,  $X_1, X_2, \dots, X_n$  are iid random variables. The distribution of  $N$  is known as the frequency distribution, and the common distribution of  $X$  is known as the severity distribution. We further assume  $N$  and  $X$  are independent. With the collective risk model, we may decompose the aggregate losses into the frequency ( $N$ ) process and the severity ( $X$ ) model. This flexibility allows the analyst to comment on these two separate components. For example, sales growth due to lower underwriting standards could lead to higher frequency of losses but might not affect severity. Similarly, inflation or other economic forces could have an impact on severity but not on frequency.

The rest of the chapter is structured as follows: Section 7.2 and Section 7.3 provide details on the individual risk model and collective risk model respectively. Section 7.4 presents methods for computing the distribution of aggregate claims. Section 7.5 discusses the effect of coverage modifications on the aggregate losses. Technical materials are summarized in Section 7.6.

---

## 7.2 Individual Risk Model

---

In this section, you learn:

- mathematical representation of the individual risk model
  - applications of individual risk model to life and non-life insurance
  - how to evaluate moments, generating functions, and the distribution function of the individual risk model
- 

### 7.2.1 Moments and Distribution

As noted earlier, for the *individual risk model*, we think of  $X_i$  as the loss from  $i$ th contract and interpret

$$S_n = X_1 + X_2 + \cdots + X_n,$$

to be the aggregate loss from all contracts in a portfolio or group of contracts. Here, the  $X_i$ 's are not necessarily identically distributed and we have

$$\mathbb{E}(S_n) = \sum_{i=1}^n \mathbb{E}(X_i).$$

Under the independence assumption on  $X_i$ 's (which implies  $\text{Cov}(X_i, X_j) = 0$  for all  $i \neq j$ ), it can further be shown that

$$\begin{aligned}\text{Var}(S_n) &= \sum_{i=1}^n \text{Var}(X_i) \\ P_{S_n}(z) &= \prod_{i=1}^n P_{X_i}(z) \\ M_{S_n}(t) &= \prod_{i=1}^n M_{X_i}(t),\end{aligned}$$

where  $P_{S_n}(\cdot)$  and  $M_{S_n}(\cdot)$  are the probability generating function (*pgf*) and the moment generating function (*mgf*) of  $S_n$ , respectively. The distribution of each  $X_i$  contains a probability mass at zero, corresponding to the event of no claims from the  $i$ th contract. One strategy to incorporate the zero mass in the distribution is to use the two-part framework:

$$X_i = I_i \times B_i = \begin{cases} 0, & \text{if } I_i = 0 \\ B_i, & \text{if } I_i = 1. \end{cases}$$

Here,  $I_i$  is a Bernoulli variable indicating whether or not a loss occurs for the  $i$ th contract, and  $B_i$  is a random variable with nonnegative support representing the amount of losses of the contract given loss occurrence. Assume that  $I_1, \dots, I_n, B_1, \dots, B_n$  are mutually independent. Denote  $\Pr(I_i = 1) = q_i$ ,  $\mu_i = \mathbb{E}(B_i)$ , and  $\sigma_i^2 = \text{Var}(B_i)$ . It can be shown (see *Technical Supplement 7.A.1* in Section 7.6 for details) that

$$\begin{aligned}\mathbb{E}(S_n) &= \sum_{i=1}^n q_i \mu_i \\ \text{Var}(S_n) &= \sum_{i=1}^n (q_i \sigma_i^2 + q_i(1-q_i)\mu_i^2) \\ P_{S_n}(z) &= \prod_{i=1}^n (1 - q_i + q_i P_{B_i}(z)) \\ M_{S_n}(t) &= \prod_{i=1}^n (1 - q_i + q_i M_{B_i}(t)).\end{aligned}$$

A special case of the above model is when  $B_i$  follows a degenerate distribution with  $\mu_i = b_i$  and  $\sigma_i^2 = 0$ . One example is term life insurance or a pure endowment insurance where  $b_i$  represents the insurance benefit amount of the  $i$ th contract.

Another strategy to accommodate the zero mass in the loss from each contract is to consider them in aggregate at the portfolio level, as in the *collective risk model*. Here, the aggregate loss is  $S_N = X_1 + \dots + X_N$ , where  $N$  is a random variable representing the number of non-zero claims that occurred out of the entire group of contracts. Thus, not every contract in the portfolio may be represented in this sum, and  $S_N = 0$  when  $N = 0$ . The collective risk model will be discussed in detail in the next section.

**Example 7.2.1. Actuarial Exam Question.** An insurance company sold 300 fire insurance policies as follows:

Number of Policies	Policy Maximum ( $M_i$ )	Probability of Claim Per Policy ( $q_i$ )
100	400	0.05
200	300	0.06

You are given:

- (i) The claim amount for each policy,  $X_i$ , is uniformly distributed between 0 and the policy maximum  $M_i$ .

- (ii) The probability of more than one claim per policy is 0.
- (iii) Claim occurrences are independent.

Calculate the mean,  $E(S_{300})$ , and variance,  $\text{Var}(S_{300})$ , of the aggregate claims. How would these results change if every claim is equal to the policy maximum?

**Example Solution.** The aggregate claims are  $S_{300} = X_1 + \dots + X_{300}$ , where  $X_1, \dots, X_{300}$  are independent but not identically distributed. Policy claims amounts are uniformly distributed on  $(0, M_i)$ , so the mean claim amount is  $M_i/2$  and the variance is  $M_i^2/12$ . Thus, for policy  $i = 1, \dots, 300$ , we have

Number of Policies	Policy Maximum ( $M_i$ )	Probability of Claim Per Policy ( $q_i$ )	Mean Amount ( $\mu_i$ )	Variance Amount ( $\sigma_i^2$ )
100	400	0.05	200	$400^2/12$
200	300	0.06	150	$300^2/12$

The mean of the aggregate claims is

$$E(S_{300}) = \sum_{i=1}^{300} q_i \mu_i = 100\{0.05(200)\} + 200\{0.06(150)\} = 2,800$$

The variance of the aggregate claims is

$$\begin{aligned} \text{Var}(S_{300}) &= \sum_{i=1}^{300} (q_i \sigma_i^2 + q_i(1 - q_i)\mu_i^2) \quad \text{since } X_i \text{'s are independent} \\ &= 100\left\{0.05\left(\frac{400^2}{12}\right) + 0.05(1 - 0.05)200^2\right\} \\ &\quad + 200\left\{0.06\left(\frac{300^2}{12}\right) + 0.06(1 - 0.06)150^2\right\} \\ &= 600,467. \end{aligned}$$

### Example 7.2.1. Continued.

Now suppose everybody receives the policy maximum  $M_i$  if a claim occurs. What is the expected aggregate loss  $E(\tilde{S})$  and variance of the aggregate loss  $\text{Var}(\tilde{S})$ ?

**Example Solution.** Each policy claim amount  $X_i$  is now deterministic and fixed at  $M_i$  instead of a randomly distributed amount, so  $\sigma_i^2 = \text{Var}(X_i) = 0$  and  $\mu_i = M_i$ . Again, the probability of a claim occurring for each policy is  $q_i$ .

Under these circumstances, the expected aggregate loss is

$$\mathbb{E}(\tilde{S}) = \sum_{i=1}^{300} q_i \mu_i = 100 \{0.05(400)\} + 200 \{0.06(300)\} = 5,600.$$

The variance of the aggregate loss is

$$\begin{aligned}\text{Var}(\tilde{S}) &= \sum_{i=1}^{300} (q_i \sigma_i^2 + q_i(1-q_i)\mu_i^2) = \sum_{i=1}^{300} (q_i(1-q_i)\mu_i^2) \\ &= 100 \{(0.05)(1-0.05)400^2\} + 200 \{(0.06)(1-0.06)300^2\} \\ &= 1,775,200.\end{aligned}$$

The individual risk model can also be used for claim frequency. If  $X_i$  denotes the number of claims from the  $i$ th contract, then  $S_n$  is interpreted as the total number of claims from the portfolio. In this case, the above two-part framework still applies since there is a probability mass at zero for contracts that do not experience any claims. Assume  $X_i$  belongs to the  $(a, b, 0)$  class with pmf denoted by  $p_{ik} = \Pr(X_i = k)$  for  $k = 0, 1, \dots$  (see Section 3.3). Let  $X_i^T$  denote the associated zero-truncated distribution in the  $(a, b, 1)$  class with pmf  $p_{ik}^T = p_{ik}/(1-p_{i0})$  for  $k = 1, 2, \dots$  (see Section 3.5.1). Using the relationship between their probability generating functions (see *Technical Supplement 7.A.2* in Section 7.6 for details):

$$P_{X_i}(z) = p_{i0} + (1-p_{i0})P_{X_i^T}(z),$$

we can write  $X_i = I_i \times B_i$  with  $q_i = \Pr(I_i = 1) = \Pr(X_i > 0) = 1 - p_{i0}$  and  $B_i = X_i^T$ . Notice that in this case, we have a zero-modified distribution since the  $I_i$  variable covers the modified probability mass at zero with  $q_i = \Pr(I_i = 1)$ , while the  $B_i = X_i^T$  covers the discrete non-zero frequency portion. See Section 3.5.1 for the relationship between zero-truncated and zero-modified distributions.

---

**Example 7.2.2.** An insurance company sold a portfolio of 100 independent homeowners insurance policies, each of which has claim frequency following a zero-modified Poisson distribution, as follows:

Type of Policy	Number of Policies	Probability of At Least 1 Claim	$\lambda$
Low-risk	40	0.03	1
High-risk	60	0.05	2

Find the expected value and variance of the claim frequency for the entire portfolio.

**Example Solution.** For each policy, we can write the zero-modified Poisson claim frequency  $N_i$  as  $N_i = I_i \times B_i$ , where

$$q_i = \Pr(I_i = 1) = \Pr(N_i > 0) = 1 - p_{i0}.$$

For the low-risk policies, we have  $q_i = 0.03$  and for the high-risk policies, we have  $q_i = 0.05$ . Further,  $B_i = N_i^T$ , the zero-truncated version of  $N_i$ . Thus, we have

$$\begin{aligned}\mu_i &= E(B_i) = E(N_i^T) = \frac{\lambda}{1 - e^{-\lambda}} \\ \sigma_i^2 &= \text{Var}(B_i) = \text{Var}(N_i^T) = \frac{\lambda[1 - (\lambda + 1)e^{-\lambda}]}{(1 - e^{-\lambda})^2}.\end{aligned}$$

Using  $n = 100$ , let the portfolio claim frequency be  $S_{100} = \sum_{i=1}^{100} N_i$ . Using the formulas above, the expected claim frequency of the portfolio is

$$\begin{aligned}E(S_{100}) &= \sum_{i=1}^{100} q_i \mu_i \\ &= 40 \left[ 0.03 \left( \frac{1}{1 - e^{-1}} \right) \right] + 60 \left[ 0.05 \left( \frac{2}{1 - e^{-2}} \right) \right] \\ &= 40(0.03)(1.5820) + 60(0.05)(2.3130) = 8.8375.\end{aligned}$$

The variance of the claim frequency of the portfolio is

$$\begin{aligned}\text{Var}(S_{100}) &= \sum_{i=1}^{100} (q_i \sigma_i^2 + q_i(1 - q_i)\mu_i^2) \\ &= 40 \left[ 0.03 \left( \frac{1 - 2e^{-1}}{(1 - e^{-1})^2} \right) + 0.03(0.97)(1.5820^2) \right] \\ &\quad + 60 \left[ 0.05 \left( \frac{2[1 - 3e^{-2}]}{(1 - e^{-2})^2} \right) + 0.05(0.95)(2.3130^2) \right] \\ &= 23.7214.\end{aligned}$$

Note that equivalently, we could have calculated the mean and variance of an individual policy directly using the relationship between the zero-modified and zero-truncated Poisson distributions (see Section 3.3).

### 7.2.2 Aggregate Loss Distribution

To understand the distribution of the aggregate loss, one could use the central limit theorem to approximate the distribution of  $S_n$  for large  $n$ . Denote  $\mu_{S_n} =$

$E(S_n)$  and  $\sigma_{S_n}^2 = \text{Var}(S_n)$  and let  $Z \sim N(0, 1)$ , a standard normal random variable with cdf  $\Phi$ . Then the cdf of  $S_n$  can be approximated as follows:

$$\begin{aligned} F_{S_n}(s) &= \Pr(S_n \leq s) = \Pr\left(\frac{S_n - \mu_{S_n}}{\sigma_{S_n}} \leq \frac{s - \mu_{S_n}}{\sigma_{S_n}}\right) \\ &\approx \Pr\left(Z \leq \frac{s - \mu_{S_n}}{\sigma_{S_n}}\right) = \Phi\left(\frac{s - \mu_{S_n}}{\sigma_{S_n}}\right). \end{aligned}$$

**Example 7.2.3. Actuarial Exam Question - Follow-Up.** As in the Example 7.2.1 earlier, an insurance company sold 300 fire insurance policies, with claim amounts  $X_i$  uniformly distributed between 0 and the policy maximum  $M_i$ . Using the normal approximation, calculate the probability that the aggregate claim amount  $S_{300}$  exceeds \$3,500.

**Example Solution.** We have seen earlier that  $E(S_{300}) = 2,800$  and  $\text{Var}(S_{300}) = 600,467$ . Then

$$\begin{aligned} \Pr(S_{300} > 3,500) &= 1 - \Pr(S_{300} \leq 3,500) \\ &\approx 1 - \Phi\left(\frac{3,500 - 2,800}{\sqrt{600,467}}\right) = 1 - \Phi(0.90334) \\ &= 1 - 0.8168 = 0.1832. \end{aligned}$$

For small  $n$ , the distribution of  $S_n$  is likely skewed, and the normal approximation would be a poor choice. To examine the aggregate loss distribution, we go back to first principles. Specifically, the distribution can be derived recursively. Define  $S_k = X_1 + \dots + X_k$ ,  $k = 1, \dots, n$ .

For  $k = 1$ :

$$F_{S_1}(s) = \Pr(S_1 \leq s) = \Pr(X_1 \leq s) = F_{X_1}(s).$$

For  $k = 2, \dots, n$ :

$$\begin{aligned} F_{S_k}(s) &= \Pr(X_1 + \dots + X_k \leq s) = \Pr(S_{k-1} + X_k \leq s) \\ &= E_{X_k} [\Pr(S_{k-1} \leq s - X_k | X_k)] = E_{X_k} [F_{S_{k-1}}(s - X_k)]. \end{aligned}$$

A special case is when  $X_i$ 's are identically distributed. Let  $F_X(x) = \Pr(X \leq x)$  be the common distribution of  $X_i$ ,  $i = 1, \dots, n$ . We define

$$F_X^{*n}(x) = \Pr(X_1 + \dots + X_n \leq x),$$

the  $n$ -fold convolution of  $F_X$ . More generally, we can compute  $F_X^{*n}$  recursively. Begin the recursion at  $k = 1$  using  $F_X^{*1}(x) = F_X(x)$ . Next, for  $k = 2$ , we have

$$\begin{aligned} F_X^{*2}(x) &= \Pr(X_1 + X_2 \leq x) = \mathbb{E}_{X_2} [\Pr(X_1 \leq x - X_2 | X_2)] \\ &= \mathbb{E}_{X_2} [F(x - X_2)] \\ &= \begin{cases} \int_0^x F(x - y) f(y) dy & \text{for continuous } X_i \text{'s} \\ \sum_{y \leq x} F(x - y) f(y) & \text{for discrete } X_i \text{'s} \end{cases} \end{aligned}$$

Recall  $F(0) = 0$ .

Similarly for  $k = n$ , we have  $S_n = X_1 + X_2 + \dots + X_n$  and

$$\begin{aligned} F^{*n}(x) &= \Pr(S_n \leq x) = \Pr(S_{n-1} + X_n \leq x) \\ &= \mathbb{E}_{X_n} [\Pr(S_{n-1} \leq x - X_n | X_n)] \\ &= \mathbb{E}_X [F^{*(n-1)}(x - X)] \\ &= \begin{cases} \int_0^x F^{*(n-1)}(x - y) f(y) dy & \text{for continuous } X_i \text{'s} \\ \sum_{y \leq x} F^{*(n-1)}(x - y) f(y) & \text{for discrete } X_i \text{'s} \end{cases} \end{aligned}$$

When the  $X_i$ 's are independent and belong to the same family of distributions, there are some simple cases where  $S_n$  has a closed form. This makes it easy to compute  $\Pr(S_n \leq x)$ . This property is known as *closed under convolution*, meaning the distribution of the sum of independent random variables belongs to the same family of distributions as that of the component variables, just with different parameters. **Table 7.1** provides a few examples.

Table 7.1. Closed Form Partial Sum Distributions

Distribution of $X_i$	Abbreviation	Distribution of $S_n$
Normal with mean $\mu_i$ and variance $\sigma_i^2$	$N(\mu_i, \sigma_i^2)$	$N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$
Exponential with mean $\theta$	$Exp(\theta)$	$Gam(n, \theta)$
Gamma with shape $\alpha_i$ and scale $\theta$	$Gam(\alpha_i, \theta)$	$Gam(\sum_{i=1}^n \alpha_i, \theta)$
Poisson with mean (and variance) $\lambda_i$	$Poi(\lambda_i)$	$Poi(\sum_{i=1}^n \lambda_i)$
Binomial with $m_i$ trials and $q$ success probability	$Bin(m_i, q)$	$Bin(\sum_{i=1}^n m_i, q)$
Geometric with mean $\beta$	$Geo(\beta)$	$NB(\beta, n)$
Negative binomial with mean $r_i\beta$ and variance $r_i\beta(1 + \beta)$	$NB(\beta, r_i)$	$NB(\beta, \sum_{i=1}^n r_i)$

**Example 7.2.4. Gamma Distribution.** Assume that  $X_1, \dots, X_n$  are independent random variables with  $X_i \sim Gam(\alpha_i, \theta)$ . The *mgf* of  $X_i$  is

$M_{X_i}(t) = (1 - \theta t)^{-\alpha_i}$ . Thus, the *mgf* of the sum  $S_n = X_1 + \dots + X_n$  is

$$\begin{aligned} M_{S_n}(t) &= \prod_{i=1}^n M_{X_i}(t) \quad \text{from the independence of } X_i \text{'s} \\ &= \prod_{i=1}^n (1 - \theta t)^{-\alpha_i} = (1 - \theta t)^{-\sum_{i=1}^n \alpha_i}, \end{aligned}$$

which is the *mgf* of a gamma random variable with parameters  $(\sum_{i=1}^n \alpha_i, \theta)$ . Thus,  $S_n \sim \text{Gam}(\sum_{i=1}^n \alpha_i, \theta)$ .

---

**Example 7.2.5. Negative Binomial Distribution.** Assume that  $X_1, \dots, X_n$  are independent random variables with  $X_i \sim NB(\beta, r_i)$ . The *pgf* of  $X_i$  is  $P_{X_i}(z) = [1 - \beta(z - 1)]^{-r_i}$ . Thus, the *pgf* of the sum  $S_n = X_1 + \dots + X_n$  is

$$\begin{aligned} P_{S_n}(z) &= E[z^{S_n}] \\ &= E[z^{X_1}] \cdots E[z^{X_n}] \quad \text{from the independence of } X_i \text{'s} \\ &= \prod_{i=1}^n P_{X_i}(z) = \prod_{i=1}^n [1 - \beta(z - 1)]^{-r_i} = [1 - \beta(z - 1)]^{-\sum_{i=1}^n r_i}, \end{aligned}$$

which is the *pgf* of a negative binomial random variable with parameters  $(\beta, \sum_{i=1}^n r_i)$ . Thus,  $S_n \sim NB(\beta, \sum_{i=1}^n r_i)$ .

---

**Example 7.2.6. Actuarial Exam Question (modified).** The annual number of doctor visits for each individual in a family of 4 has geometric distribution with mean 1.5. The annual numbers of visits for the family members are mutually independent. An insurance pays 100 per doctor visit beginning with the 4th visit per family. Calculate the probability that the family will not receive an insurance payment this year.

**Example Solution.** Let  $X_i \sim Geo(\beta = 1.5)$  be the number of doctor visits for one individual in the family and  $S_4 = X_1 + X_2 + X_3 + X_4$  be the number of doctor visits for the family. The sum of 4 independent geometric random variables each with mean  $\beta = 1.5$  follows a negative binomial distribution, i.e.  $S_4 \sim NB(\beta = 1.5, r = 4)$ .

If the insurance pays 100 per visit beginning with the 4th visit for the family, then the family will not receive an insurance payment if they have less than 4

claims. This probability is

$$\begin{aligned}\Pr(S_4 < 4) &= \Pr(S_4 = 0) + \Pr(S_4 = 1) + \Pr(S_4 = 2) + \Pr(S_4 = 3) \\ &= (1 + 1.5)^{-4} + \frac{4(1.5)}{(1 + 1.5)^5} + \frac{4(5)(1.5^2)}{2(1 + 1.5)^6} + \frac{4(5)(6)(1.5^3)}{3!(1 + 1.5)^7} \\ &= 0.0256 + 0.0614 + 0.0922 + 0.1106 = 0.2898.\end{aligned}$$

## 7.3 Collective Risk Model

In this section, you learn:

- mathematical representation of the collective risk model
- how to evaluate moments, generating functions, and the distribution function of the collective risk model
- applications of collective risk model in stop-loss insurance
- Tweedie compound Poisson distribution as a special case of the collective risk model

### 7.3.1 Moments and Distribution

Under the collective risk model  $S_N = X_1 + \dots + X_N$ ,  $\{X_i\}$  are *iid*, and independent of  $N$ . Let  $\mu = E(X_i)$  and  $\sigma^2 = \text{Var}(X_i)$  for all  $i$ . Thus, conditional on  $N$ , we have that the expectation of the sum is the sum of expectations and that the variance of the sum is the sum of variances,

$$\begin{aligned}E(S|N) &= E(X_1 + \dots + X_N|N) = \mu N \\ \text{Var}(S|N) &= \text{Var}(X_1 + \dots + X_N|N) = \sigma^2 N.\end{aligned}$$

Using the law of iterated expectations from Appendix Section ??, the mean of the aggregate loss is

$$E(S_N) = E_N[E_S(S|N)] = E_N(N\mu) = \mu E(N).$$

Using the law of total variance from Appendix Section ??, the variance of the aggregate loss is

$$\begin{aligned}\text{Var}(S_N) &= E_N[\text{Var}(S_N|N)] + \text{Var}_N[E(S_N|N)] \\ &= E_N[\sigma^2 N] + \text{Var}_N[\mu N] \\ &= \sigma^2 E[N] + \mu^2 \text{Var}[N].\end{aligned}$$

**Special Case: Poisson Distributed Frequency.** If  $N \sim Poi(\lambda)$ , then

$$\begin{aligned} E(N) &= \text{Var}(N) = \lambda \\ E(S_N) &= \lambda E(X) \\ \text{Var}(S_N) &= \lambda(\sigma^2 + \mu^2) = \lambda E(X^2). \end{aligned}$$


---

**Example 7.3.1. Actuarial Exam Question.** The number of accidents follows a Poisson distribution with mean 12. Each accident generates 1, 2, or 3 claimants with probabilities  $1/2$ ,  $1/3$ , and  $1/6$  respectively.

Calculate the variance in the total number of claimants.

**Example Solution.**

$$\begin{aligned} E(X^2) &= 1^2 \left( \frac{1}{2} \right) + 2^2 \left( \frac{1}{3} \right) + 3^2 \left( \frac{1}{6} \right) = \frac{10}{3} \\ \Rightarrow \text{Var}(S_N) &= \lambda E(X^2) = 12 \left( \frac{10}{3} \right) = 40. \end{aligned}$$

Alternatively, using the general approach,  $\text{Var}(S_N) = \sigma^2 E(N) + \mu^2 \text{Var}(N)$ , where

$$\begin{aligned} E(N) &= \text{Var}(N) = 12 \\ \mu &= E(X) = 1 \left( \frac{1}{2} \right) + 2 \left( \frac{1}{3} \right) + 3 \left( \frac{1}{6} \right) = \frac{5}{3} \\ \sigma^2 &= E(X^2) - [E(X)]^2 = \frac{10}{3} - \frac{25}{9} = \frac{5}{9} \\ \Rightarrow \text{Var}(S_N) &= \left( \frac{5}{9} \right) (12) + \left( \frac{5}{3} \right)^2 (12) = 40. \end{aligned}$$

In general, the moments of  $S_N$  can be derived from its moment generating function (*mgf*). Because  $X_i$ 's are *iid*, we denote the *mgf* of  $X$  as  $M_X(t) = E(e^{tX})$ . Using the law of iterated expectations, the *mgf* of  $S_N$  is

$$\begin{aligned} M_{S_N}(t) &= E(e^{tS_N}) = E_N[ E(e^{tS_N}|N) ] \\ &= E_N[ E(e^{t(X_1+\dots+X_N)}) ] = E_N[ E(e^{tX_1}) \dots E(e^{tX_N}) ] \quad \text{since } X_i \text{'s are independent} \\ &= E_N[ (M_X(t))^N ]. \end{aligned}$$

Now, recall that the probability generating function (*pgf*) of  $N$  is  $P_N(z) = E(z^N)$ . Denote  $M_X(t) = z$ . Substituting into the expression for the *mgf* of  $S_N$  above, it is shown

$$M_{S_N}(t) = E(z^N) = P_N(z) = P_N[M_X(t)].$$

Similarly, if  $S_N$  is discrete, one can show the *pgf* of  $S_N$  is:

$$P_{S_N}(z) = P_N[P_X(z)].$$

To get  $E(S_N) = M'_{S_N}(0)$ , we use the chain rule

$$M'_{S_N}(t) = \frac{\partial}{\partial t} P_N(M_X(t)) = P'_N(M_X(t))M'_X(t)$$

and recall  $M_X(0) = 1, M'_X(0) = E(X) = \mu, P'_N(1) = E(N)$ . So,

$$E(S_N) = M'_{S_N}(0) = P'_N(M_X(0))M'_X(0) = \mu E(N).$$

Similarly, one could use relation  $E(S_N^2) = M''_{S_N}(0)$  to get

$$\text{Var}(S_N) = \sigma^2 E(N) + \mu^2 \text{Var}(N).$$

**Special Case. Poisson Frequency.** Let  $N \sim Poi(\lambda)$ . Thus, the *pgf* of  $N$  is  $P_N(z) = e^{\lambda(z-1)}$  and the *mgf* of  $S_N$  is

$$M_{S_N}(t) = P_N[M_X(t)] = e^{\lambda(M_X(t)-1)}.$$

Taking derivatives yields

$$\begin{aligned} M'_{S_N}(t) &= e^{\lambda(M_X(t)-1)} \lambda M'_X(t) = M_{S_N}(t) \lambda M'_X(t) \\ M''_{S_N}(t) &= M_{S_N}(t) \lambda M''_X(t) + [M_{S_N}(t) \lambda M'_X(t)] \lambda M'_X(t). \end{aligned}$$

Evaluating these at  $t = 0$  yields

$$E(S_N) = M'_{S_N}(0) = \lambda E(X) = \lambda \mu$$

and

$$\begin{aligned} M''_{S_N}(0) &= \lambda E(X^2) + \lambda^2 \mu^2 \\ \Rightarrow \text{Var}(S_N) &= \lambda E(X^2) + \lambda^2 \mu^2 - (\lambda \mu)^2 = \lambda E(X^2). \end{aligned}$$


---

**Example 7.3.2. Actuarial Exam Question.** You are the producer of a television quiz show that gives cash prizes. The number of prizes,  $N$ , and prize amount,  $X$ , have the following distributions:

$n$	$\Pr(N = n)$	$x$	$\Pr(X = x)$
1	0.8	0	0.2
2	0.2	100	0.7
		1000	0.1

Your budget for prizes equals the expected aggregate cash prizes plus the standard deviation of aggregate cash prizes. Calculate your budget.

**Example Solution.** We need to calculate the mean and standard deviation of the aggregate (sum) of cash prizes. The moments of the frequency distribution  $N$  are

$$\begin{aligned} E(N) &= 1(0.8) + 2(0.2) = 1.2 \\ E(N^2) &= 1^2(0.8) + 2^2(0.2) = 1.6 \\ \text{Var}(N) &= E(N^2) - [E(N)]^2 = 0.16. \end{aligned}$$

The moments of the severity distribution  $X$  are

$$\begin{aligned} E(X) &= 0(0.2) + 100(0.7) + 1000(0.1) = 170 = \mu \\ E(X^2) &= 0^2(0.2) + 100^2(0.7) + 1000^2(0.1) = 107,000 \\ \text{Var}(X) &= E(X^2) - [E(X)]^2 = 78,100 = \sigma^2. \end{aligned}$$

Thus, the mean and variance of the aggregate cash prize are

$$\begin{aligned} E(S_N) &= \mu E(N) = 170(1.2) = 204 \\ \text{Var}(S_N) &= \sigma^2 E(N) + \mu^2 \text{Var}(N) \\ &= 78,100(1.2) + 170^2(0.16) = 98,344. \end{aligned}$$

This gives the following required budget

$$\begin{aligned} \text{Budget} &= E(S_N) + \sqrt{\text{Var}(S_N)} \\ &= 204 + \sqrt{98,344} = 517.60. \end{aligned}$$

---

The distribution of  $S_N$  is called a compound distribution, and it can be derived based on the convolution of  $F_X$  as follows:

$$\begin{aligned} F_{S_N}(s) &= \Pr(X_1 + \dots + X_N \leq s) \\ &= E[\Pr(X_1 + \dots + X_N \leq s | N = n)] \\ &= E[F_X^{*n}(s)] \\ &= p_0 + \sum_{n=1}^{\infty} p_n F_X^{*n}(s). \end{aligned}$$


---

**Example 7.3.3. Actuarial Exam Question.** The number of claims in a period has a geometric distribution with mean 4. The amount of each claim  $X$  follows  $\Pr(X = x) = 0.25$ ,  $x = 1, 2, 3, 4$ , i.e. a discrete uniform distribution on  $\{1, 2, 3, 4\}$ . The number of claims and the claim amounts are independent. Let  $S_N$  denote the aggregate claim amount in the period. Calculate  $F_{S_N}(3)$ .

**Example Solution.** By definition, we have

$$\begin{aligned}
 F_{S_N}(3) &= \Pr\left(\sum_{i=1}^N X_i \leq 3\right) = \sum_{n=0}^{\infty} \Pr\left(\sum_{i=1}^n X_i \leq 3 | N = n\right) \Pr(N = n) \\
 &= \sum_n F^{*n}(3) p_n = \sum_{n=0}^3 F^{*n}(3) p_n \\
 &= p_0 + F^{*1}(3) p_1 + F^{*2}(3) p_2 + F^{*3}(3) p_3.
 \end{aligned}$$

Because  $N \sim \text{Geo}(\beta = 4)$ , we know that

$$p_n = \frac{1}{1+\beta} \left( \frac{\beta}{1+\beta} \right)^n = \frac{1}{5} \left( \frac{4}{5} \right)^n.$$

For the claim severity distribution, recursively, we have

$$\begin{aligned}
 F^{*1}(3) &= \Pr(X \leq 3) = \frac{3}{4} \\
 F^{*2}(3) &= \sum_{y \leq 3} F^{*1}(3-y)f(y) = F^{*1}(2)f(1) + F^{*1}(1)f(2) \\
 &= \frac{1}{4} [F^{*1}(2) + F^{*1}(1)] = \frac{1}{4} [\Pr(X \leq 2) + \Pr(X \leq 1)] \\
 &= \frac{1}{4} \left( \frac{2}{4} + \frac{1}{4} \right) = \frac{3}{16} \\
 F^{*3}(3) &= \Pr(X_1 + X_2 + X_3 \leq 3) = \Pr(X_1 = X_2 = X_3 = 1) = \left( \frac{1}{4} \right)^3.
 \end{aligned}$$

Notice that we did not need to recursively calculate  $F^{*3}(3)$  by recognizing that each  $X \in \{1, 2, 3, 4\}$ , so the only way of obtaining  $X_1 + X_2 + X_3 \leq 3$  is to have  $X_1 = X_2 = X_3 = 1$ . Additionally, for  $n \geq 4$ ,  $F^{*n}(3) = 0$  since it is impossible for the sum of 4 or more  $X$ 's to be less than 3. For  $n = 0$ ,  $F^{*0}(3) = 1$  since the sum of 0  $X$ 's is 0, which is always less than 3. Laying out the probabilities systematically,

$x$	$F^{*1}(x)$	$F^{*2}(x)$	$F^{*3}(x)$
0			
1	$\frac{1}{4}$	0	
2	$\frac{2}{4}$	$\left(\frac{1}{4}\right)^2$	
3	$\frac{3}{4}$	$\frac{3}{16}$	$\left(\frac{1}{4}\right)^3$

Finally,

$$\begin{aligned} F_{S_N}(3) &= p_0 + F^{*1}(3) p_1 + F^{*2}(3) p_2 + F^{*3}(3) p_3 \\ &= \frac{1}{5} + \frac{3}{4} \left( \frac{4}{25} \right) + \frac{3}{16} \left( \frac{16}{125} \right) + \frac{1}{64} \left( \frac{64}{625} \right) = 0.3456. \end{aligned}$$

**Example 7.3.4. Convolution Method to Compute the Aggregate Loss Distribution.** Consider the Wisconsin Property Fund data that was introduced in Section 1.3 and is available in Appendix Section ???. Specifically, we examine building and content claims with frequency of claims given by the variable `Freq` and amount of claims given by `BCClaim`. Assume a Poisson distribution for the frequency and a gamma distribution for the severity. The following block of R code illustrates how to retrieve the data and reviews parameter estimation from prior chapters.

```
datraw <- read.csv("Data/WiscPropFund.csv")
# remove extreme observations to speed up the evaluation of distribution of
# aggregate losses
index <- which(datraw$Freq < 100 & datraw$BCClaim < 250000)
dat <- datraw[index, ]
# head(dat, n=3) tail(dat, n=3)

# Assume a Poisson for claim frequency
lambda <- mean(dat$Freq)
# print(lambda) Assume a gamma for claim severity
index <- which(dat$BCClaim > 0)
n <- dat$Freq[index]
xbar <- dat$BCClaim[index]/dat$Freq[index]
fit <- glm(xbar ~ 1, family = Gamma(link = "log"), weight = 1/n)
mu <- unname(exp(fit$coefficients))
phi <- summary(fit)$dispersion
a = 1/phi
s = mu * phi
# print(c(a, s))
```

With the parameter estimates in place, we are now in a position to calculate distribution of  $S = X_1 + X_2 + \dots + X_N$  using convolution method. Figure 7.1 summarizes the aggregate loss distribution. The following block of code demonstrates its calculation.

```
Nmax <- 1000
# CDF
FAGg <- function(y) {
  re <- dpois(0, lambda)
  for (i in 1:Nmax) {
    re <- re + dpois(i, lambda) * pgamma(y, shape = i * a, scale = s)
  }
  re <- ifelse(y < 0, NA, re)
```

```

        return(re)
    }
    # PDF
    fAgg <- function(y) {
        re <- dpois(0, lambda)
        for (i in 1:Nmax) {
            re <- re + dpois(i, lambda) * dgamma(y * (y > 0) - 1 * (y <= 0), shape = i *
                a, scale = s)
        }
        re <- ifelse(y < 0, NA, re)
        return(re)
    }
    # Numerical examples
    obs <- c(-1, 0, 1, 10, 100, 1000, 10000, 100000, 1000000)
    # FAgg(obs) fAgg(obs)
}

```

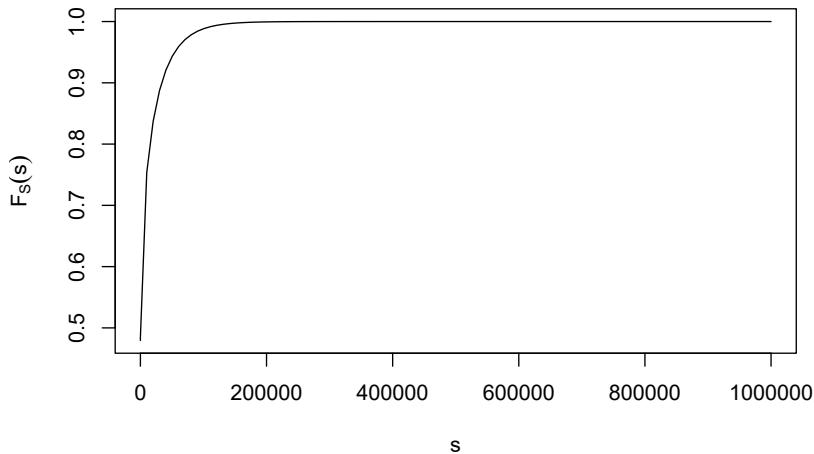


FIGURE 7.1: Aggregate Loss Distribution for Wisconsin Property Fund Building and Loss Claims

---

When  $E(N)$  and  $\text{Var}(N)$  are known, one may also use a type of central limit theorem to approximate the distribution of  $S_N$  as in the individual risk model. That is,  $\frac{S_N - E(S_N)}{\sqrt{\text{Var}(S_N)}}$  approximately follows the standard normal distribution  $N(0, 1)$ . From this type of central limit theorem, the approximation works well if  $E[N]$  is sufficiently large.

---

**Example 7.3.5. Actuarial Exam Question.** You are given:

	Mean	Standard Deviation
Number of Claims	8	3
Individual Losses	10,000	3,937

As a benchmark, use the normal approximation to determine the probability that the aggregate loss will exceed 150% of the expected loss.

**Example Solution.** To use the normal approximation, we must first find the mean and variance of the aggregate loss  $S$

$$\begin{aligned} E(S_N) &= \mu E(N) = 10,000(8) = 80,000 \\ \text{Var}(S_N) &= \sigma^2 E(N) + \mu^2 \text{Var}(N) \\ &= 3937^2(8) + 10000^2(3^2) = 1,023,999,752 \\ \sqrt{\text{Var}(S_N)} &= 31,999.996 \approx 32,000. \end{aligned}$$

Then under the normal approximation, aggregate loss  $S_N$  is approximately normal with mean 80,000 and standard deviation 32,000. The probability that  $S_N$  will exceed 150% of the expected aggregate loss is therefore

$$\begin{aligned} \Pr(S_N > 1.5E(S_N)) &= \Pr\left(\frac{S_N - E(S_N)}{\sqrt{\text{Var}(S_N)}} > \frac{1.5 E(S_N) - E(S_N)}{\sqrt{\text{Var}(S_N)}}\right) \\ &\approx \Pr\left(Z > \frac{0.5 E(S_N)}{\sqrt{\text{Var}(S_N)}}\right), \quad \text{where } Z \sim N(0, 1) \\ &= \Pr\left(Z > \frac{0.5(80,000)}{32,000}\right) = \Pr(Z > 1.25) \\ &= 1 - \Phi(1.25) = 0.1056. \end{aligned}$$

**Example 7.3.6. Actuarial Exam Question.** For an individual over 65:

- (i) The number of pharmacy claims is a Poisson random variable with mean 27.
- (ii) The amount of each pharmacy claim is uniformly distributed between 5 and 97.
- (iii) The amounts of the claims and the number of claims are mutually independent.

Estimate the probability that aggregate claims for this individual will exceed 2000 using the normal approximation.

**Example Solution.** We have claim frequency  $N \sim Poi(\lambda = 25)$  and claim severity  $X \sim U(5, 95)$ . To use the normal approximation, we need to find the mean and variance of the aggregate claims  $S_N$ . Note

$$\begin{aligned} E(N) &= 25 & \text{Var}(N) &= 25 \\ E(X) &= \frac{5+95}{2} = 50 = \mu & \text{Var}(X) &= \frac{(95-5)^2}{12} = 675 = \sigma^2. \end{aligned}$$

Then for  $S_N$ ,

$$\begin{aligned} E(S_N) &= \mu E(N) = 50(25) = 1,250 \\ \text{Var}(S_N) &= \sigma^2 E(N) + \mu^2 \text{Var}(N) \\ &= 675(25) + 50^2(25) = 79,375. \end{aligned}$$

Using the normal approximation,  $S_N$  is approximately normal with mean 1,250 and variance 79,375. The probability that  $S_N$  exceeds 2,000 is

$$\begin{aligned} \Pr(S_N > 2,000) &= \Pr\left(\frac{S_N - E(S_N)}{\sqrt{\text{Var}(S_N)}} > \frac{2,000 - E(S_N)}{\sqrt{\text{Var}(S_N)}}\right) \\ &\approx \Pr\left(Z > \frac{2,000 - 1,250}{\sqrt{79,375}}\right), \quad \text{where } Z \sim N(0, 1) \\ &= \Pr(Z > 2.662) = 1 - \Phi(2.662) = 0.003884. \end{aligned}$$

### 7.3.2 Stop-loss Insurance

Recall the coverage modifications on the individual policy level in Section 5.1. Insurance on the aggregate loss  $S_N$ , subject to a deductible  $d$ , is called *net stop-loss insurance*. The expected value of the amount of the aggregate loss in excess of the deductible,

$$E[(S_N - d)_+]$$

is known as the *net stop-loss premium*.

To calculate the net stop-loss premium, we have

$$\begin{aligned} E(S_N - d)_+ &= \begin{cases} \int_d^\infty (s - d) f_{S_N}(s) ds & \text{for continuous } S_N \\ \sum_{s>d} (s - d) f_{S_N}(s) & \text{for discrete } S_N \end{cases} \\ &= E(S_N) - E(S_N \wedge d) \end{aligned}$$

**Example 7.3.7. Actuarial Exam Question.** In a given week, the number of projects that require you to work overtime has a geometric distribution

with  $\beta = 2$ . For each project, the distribution of the number of overtime hours in the week,  $X$ , is as follows:

$x$	$f(x)$
5	0.2
10	0.3
20	0.5

The number of projects and the number of overtime hours are independent. You will get paid for overtime hours in excess of 15 hours in the week. Calculate the expected number of overtime hours for which you will get paid in the week.

**Example Solution.** The number of projects in a week requiring overtime work has distribution  $N \sim Geo(\beta = 2)$ , while the number of overtime hours worked per project has distribution  $X$  as described above. The aggregate number of overtime hours in a week is  $S_N$  and we are therefore looking for

$$E(S_N - 15)_+ = E(S_N) - E(S_N \wedge 15).$$

To find  $E(S_N) = E(X) E(N)$ , we have

$$\begin{aligned} E(X) &= 5(0.2) + 10(0.3) + 20(0.5) = 14 \\ E(N) &= 2 \\ \Rightarrow E(S) &= E(X) E(N) = 14(2) = 28. \end{aligned}$$

To find  $E(S_N \wedge 15) = 0 \Pr(S_N = 0) + 5 \Pr(S_N = 5) + 10 \Pr(S_N = 10) + 15 \Pr(S_N \geq 15)$ , we have

$$\begin{aligned} \Pr(S_N = 0) &= \Pr(N = 0) = \frac{1}{1 + \beta} = \frac{1}{3} \\ \Pr(S_N = 5) &= \Pr(X = 5, N = 1) = 0.2 \left( \frac{2}{9} \right) = \frac{0.4}{9} \\ \Pr(S_N = 10) &= \Pr(X = 10, N = 1) + \Pr(X_1 = X_2 = 5, N = 2) \\ &= 0.3 \left( \frac{2}{9} \right) + (0.2)(0.2) \left( \frac{4}{27} \right) = 0.0726 \\ \Pr(S_N \geq 15) &= 1 - \left( \frac{1}{3} + \frac{0.4}{9} + 0.0726 \right) = 0.5496 \\ \Rightarrow E(S_N \wedge 15) &= 0 \Pr(S_N = 0) + 5 \Pr(S_N = 5) + 10 \Pr(S_N = 10) + 15 \Pr(S_N \geq 15) \\ &= 0 \left( \frac{1}{3} \right) + 5 \left( \frac{0.4}{9} \right) + 10(0.0726) + 15(0.5496) = 9.193. \end{aligned}$$

Therefore,

$$\begin{aligned} E(S_N - 15)_+ &= E(S_N) - E(S_N \wedge 15) \\ &= 28 - 9.193 = 18.807. \end{aligned}$$

---

**Recursive Net Stop-Loss Premium Calculation.** For the discrete case, this can be computed recursively as

$$\mathbb{E} [(S_N - (j+1)h)_+] = \mathbb{E} [(S_N - jh)_+] - h [1 - F_{S_N}(jh)] .$$

This assumes that the support of  $S_N$  is equally spaced over units of  $h$ .

To establish this, we assume that  $h = 1$ . We have

$$\begin{aligned}\mathbb{E} [(S_N - (j+1))_+] &= \mathbb{E}(S_N) - \mathbb{E}[S_N \wedge (j+1)] , \quad \text{and} \\ \mathbb{E} [(S_N - j)_+] &= \mathbb{E}(S_N) - \mathbb{E}[S_N \wedge j]\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E} [(S_N - (j+1))_+] - \mathbb{E} [(S_N - j)_+] &= \{\mathbb{E}(S_N) - \mathbb{E}(S_N \wedge (j+1))\} - \{\mathbb{E}(S_N) - \mathbb{E}(S_N \wedge j)\} \\ &= \mathbb{E}(S_N \wedge j) - \mathbb{E}[S \wedge (j+1)]\end{aligned}$$

We can write

$$\begin{aligned}\mathbb{E}[S_N \wedge (j+1)] &= \sum_{x=0}^j xf_{S_N}(x) + (j+1) \Pr(S_N \geq j+1) \\ &= \sum_{x=0}^{j-1} xf_{S_N}(x) + j \Pr(S_N = j) + (j+1) \Pr(S_N \geq j+1)\end{aligned}$$

Similarly,

$$\mathbb{E}(S_N \wedge j) = \sum_{x=0}^{j-1} xf_{S_N}(x) + j \Pr(S_N \geq j)$$

With these expressions, we have

$$\begin{aligned}\mathbb{E} [(S_N - (j+1))_+] - \mathbb{E} [(S_N - j)_+] &= \mathbb{E}(S_N \wedge j) - \mathbb{E}[S \wedge (j+1)] \\ &= \left\{ \sum_{x=0}^{j-1} xf_{S_N}(x) + j \Pr(S_N \geq j) \right\} - \left\{ \sum_{x=0}^{j-1} xf_{S_N}(x) + j \Pr(S_N = j) + (j+1) \Pr(S_N \geq j+1) \right\} \\ &= j [\Pr(S_N \geq j) - \Pr(S_N = j)] - (j+1) \Pr(S_N \geq j+1) \\ &= j \Pr(S_N > j) - (j+1) \Pr(S_N \geq j+1) \quad (\text{note } \Pr(S_N > j) = \Pr(S_N \geq j+1)) \\ &= -\Pr(S_N \geq j+1) = -[1 - F_{S_N}(j)] ,\end{aligned}$$

as required.

**Example 7.3.8. Actuarial Exam Question - Continued.** Recall that the goal of this question was to calculate  $E(S_N - 15)_+$ . Note that the support of  $S_N$  is equally spaced over units of 5, so this question can also be done recursively, using the expression above with steps of  $h = 5$ :

- Step 1:

$$\begin{aligned} E(S_N - 5)_+ &= E(S_N) - 5[1 - \Pr(S_N \leq 0)] \\ &= 28 - 5 \left(1 - \frac{1}{3}\right) = \frac{74}{3} = 24.6667. \end{aligned}$$

- Step 2:

$$\begin{aligned} E(S_N - 10)_+ &= E(S_N - 5)_+ - 5[1 - \Pr(S_N \leq 5)] \\ &= \frac{74}{3} - 5 \left(1 - \frac{1}{3} - \frac{0.4}{9}\right) = 21.555. \end{aligned}$$

- Step 3:

$$\begin{aligned} E(S_N - 15)_+ &= E(S_N - 10)_+ - 5[1 - \Pr(S_N \leq 10)] \\ &= E(S_N - 10)_+ - 5\Pr(S_N \geq 15) \\ &= 21.555 - 5(0.5496) = 18.807. \end{aligned}$$


---

### 7.3.3 Closed-form Distributions

There are a few combinations of claim frequency and severity distributions that result in an easy-to-compute distribution for aggregate losses. This section provides some simple examples. Although these examples are computationally convenient, they are generally too simple to be used in practice.

**Example 7.3.9. Geometric Frequency, Exponential Severity.** One has a closed-form expression for the aggregate loss distribution by assuming a geometric frequency distribution and an exponential severity distribution.

Assume that claim count  $N$  is geometric with mean  $E(N) = \beta$ , and that claim amount  $X$  is exponential with  $E(X) = \theta$ . Recall that the *pgf* of  $N$  and the *mgf* of  $X$  are:

$$\begin{aligned} P_N(z) &= \frac{1}{1 - \beta(z - 1)} \\ M_X(t) &= \frac{1}{1 - \theta t}. \end{aligned}$$

Thus, the *mgf* of aggregate loss  $S_N$  can be expressed two ways (for details, see *Technical Supplement 7.A.3*)

$$\begin{aligned} M_{S_N}(t) &= P_N[M_X(t)] = \frac{1}{1 - \beta(\frac{1}{1-\theta t} - 1)} \\ &= 1 + \frac{\beta}{1+\beta} ([1 - \theta(1+\beta)t]^{-1} - 1) \end{aligned} \quad (7.1)$$

$$= \frac{1}{1+\beta}(1) + \frac{\beta}{1+\beta} \left( \frac{1}{1 - \theta(1+\beta)t} \right). \quad (7.2)$$

From (7.1), we note that  $S_N$  is equivalent to the compound distribution of  $S_N = X_1^* + \dots + X_{N^*}^*$ , where  $N^*$  is a Bernoulli with mean  $\beta/(1+\beta)$  and  $X^*$  is an exponential with mean  $\theta(1+\beta)$ . To see this, we examine the *mgf* of  $S$ :

$$M_{S_N}(t) = P_N[M_X(t)] = P_{N^*}[M_{X^*}(t)],$$

where

$$\begin{aligned} P_{N^*}(z) &= 1 + \frac{\beta}{1+\beta}(z-1), \\ M_{X^*}(t) &= \frac{1}{1 - \theta(1+\beta)t}. \end{aligned}$$

From (7.2), we note that  $S_N$  is also equivalent to a two-point mixture of 0 and  $X^*$ . Specifically,

$$S_N = \begin{cases} 0 & \text{with probability } \Pr(N^* = 0) = 1/(1+\beta) \\ X^* & \text{with probability } \Pr(N^* = 1) = \beta/(1+\beta). \end{cases}$$

The distribution function of  $S_N$  is:

$$\begin{aligned} \Pr(S_N = 0) &= \frac{1}{1+\beta} \\ \Pr(S_N > s) &= \Pr(X^* > s) = \frac{\beta}{1+\beta} \exp\left(-\frac{s}{\theta(1+\beta)}\right) \end{aligned}$$

with pdf for  $s > 0$ ,

$$f_{S_N}(s) = \frac{\beta}{\theta(1+\beta)^2} \exp\left(-\frac{s}{\theta(1+\beta)}\right).$$

**Example 7.3.10. Exponential Severity.** Consider a collective risk model with an exponential severity and an arbitrary frequency distribution. Recall

that if  $X_i \sim Exp(\theta)$ , then the sum of *iid* exponential random variables,  $S_n = X_1 + \dots + X_n$ , has a gamma distribution, i.e.  $S_n \sim Gam(n, \theta)$ . This has cdf:

$$\begin{aligned} F_X^{*n}(s) &= \Pr(S_n \leq s) = \int_0^s \frac{1}{\Gamma(n)\theta^n} s^{n-1} \exp\left(-\frac{s}{\theta}\right) ds \\ &= 1 - \sum_{j=0}^{n-1} \frac{1}{j!} \left(\frac{s}{\theta}\right)^j e^{-s/\theta}. \end{aligned}$$

The last equality is derived by applying integration by parts  $n - 1$  times.

For the aggregate loss distribution, we can interchange the order of summations in the second line below to get

$$\begin{aligned} F_S(s) &= p_0 + \sum_{n=1}^{\infty} p_n F_X^{*n}(s) \\ &= 1 - \sum_{n=1}^{\infty} p_n \sum_{j=0}^{n-1} \frac{1}{j!} \left(\frac{s}{\theta}\right)^j e^{-s/\theta} \\ &= 1 - e^{-s/\theta} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{s}{\theta}\right)^j \bar{P}_j \end{aligned}$$

where  $\bar{P}_j = p_{j+1} + p_{j+2} + \dots = \Pr(N > j)$  is the “survival function” of the claims count distribution.

---

### 7.3.4 Tweedie Distribution

In this section, we examine a particular compound distribution where the number of claims has a Poisson distribution and the amount of claims has a gamma distribution. This specification leads to what is known as a Tweedie distribution. The Tweedie distribution has a mass probability at zero and a continuous component for positive values. Because of this feature, it is widely used in insurance claims modeling, where the zero mass is interpreted as no claims and the positive component as the amount of claims.

Specifically, consider the collective risk model  $S_N = X_1 + \dots + X_N$ . Suppose that  $N$  has a Poisson distribution with mean  $\lambda$ , and each  $X_i$  has a gamma distribution with shape parameter  $\alpha$  and scale parameter  $\gamma$ . The Tweedie distribution is derived as the Poisson sum of gamma variables. To understand the distribution of  $S_N$ , we first examine the mass probability at zero. The aggregate loss is zero when no claims occurred, i.e.

$$\Pr(S_N = 0) = \Pr(N = 0) = e^{-\lambda}.$$

In addition, note that  $S_N$  conditional on  $N = n$ , denoted by  $S_n = X_1 + \dots + X_n$ , follows a gamma distribution with shape  $n\alpha$  and scale  $\gamma$ . Thus, for  $s > 0$ , the

density of a Tweedie distribution can be calculated as

$$\begin{aligned} f_{S_N}(s) &= \sum_{n=1}^{\infty} p_n f_{S_n}(s) \\ &= \sum_{n=1}^{\infty} e^{-\lambda} \frac{(\lambda)^n}{n!} \frac{\gamma^{na}}{\Gamma(n\alpha)} s^{n\alpha-1} e^{-s\gamma}. \end{aligned}$$

Thus, the Tweedie distribution can be thought of a mixture of zero and a positive valued distribution, which makes it a convenient tool for modeling insurance claims and for calculating pure premiums. The mean and variance of the Tweedie compound Poisson model are:

$$E(S_N) = \lambda \frac{\alpha}{\gamma} \quad \text{and} \quad \text{Var}(S_N) = \lambda \frac{\alpha(1+\alpha)}{\gamma^2}.$$

As another important feature, the Tweedie distribution is a special case of exponential dispersion models, a class of models used to describe the random component in generalized linear models. To see this, we consider the following reparameterization:

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \quad \alpha = \frac{2-p}{p-1}, \quad \frac{1}{\gamma} = \phi(p-1)\mu^{p-1}.$$

With the above relationships, one can show that the distribution of  $S_N$  is

$$f_{S_N}(s) = \exp \left[ \frac{1}{\phi} \left( \frac{-s}{(p-1)\mu^{p-1}} - \frac{\mu^{2-p}}{2-p} \right) + C(s; \phi) \right]$$

where

$$C(s; \phi) = \begin{cases} 0 & \text{if } s = 0 \\ \log \sum_{n \geq 1} \left\{ \frac{(1/\phi)^{1/(p-1)} s^{(2-p)/(p-1)}}{(2-p)(p-1)^{(2-p)/(p-1)}} \right\}^n \frac{1}{n! \Gamma[n(2-p)/(p-1)]s} & \text{if } s > 0. \end{cases}$$

Hence, the distribution of  $S_N$  belongs to the exponential family with parameters  $\mu$ ,  $\phi$ , and  $1 < p < 2$ , and we have

$$E(S_N) = \mu \quad \text{and} \quad \text{Var}(S_N) = \phi\mu^p.$$

This allows us to use the Tweedie distribution with generalized linear models to model claims. It is also worth mentioning the two limiting cases of the Tweedie model:  $p \rightarrow 1$  results in the Poisson distribution and  $p \rightarrow 2$  results in the gamma distribution. Thus, the Tweedie model accommodates the situations in between the gamma and Poisson distributions, which makes intuitive sense as it is the Poisson sum of gamma random variables.

## 7.4 Computing the Aggregate Claims Distribution

In this section, you learn:

- the recursive method to compute the aggregate claims distribution
- the simulation approach to compute the aggregate claims distribution

Computing the distribution of aggregate losses is a difficult, yet important, problem. As we have seen, for both individual risk model and collective risk model, computing the distribution frequently involves the evaluation of a  $n$ -fold convolution. To make the problem tractable, one strategy is to use a distribution that is easy to evaluate to approximate the aggregate loss distribution. For instance, normal distribution is a natural choice based on central limit theorem where parameters of the normal distribution can be estimated by matching the moments. This approach has its strength and limitations. Its main advantage is the ease of computation. The disadvantages are: first, the size and direction of approximation error are unknown; second, the approximation may fail to capture some special features of the aggregate loss such as mass point at zero.

This section discusses two practical approaches to computing the distribution of aggregate loss, the recursive method and simulation.

### 7.4.1 Recursive Method

The recursive method applies to compound models where the frequency component  $N$  belongs to either  $(a, b, 0)$  or  $(a, b, 1)$  class (see Sections 3.3 and 3.5.1) and the severity component  $X$  has a discrete distribution. For continuous  $X$ , a common practice is to first discretize the severity distribution, after which the recursive method is ready to apply.

Assume that  $N$  is in the  $(a, b, 1)$  class so that  $p_k = (a + \frac{b}{k}) p_{k-1}$ ,  $k = 2, 3, \dots$ . Further assume that the support of  $X$  is  $\{0, 1, \dots, m\}$ , discrete and finite. Then, the probability function of  $S_N$  is:

$$\begin{aligned} f_{S_N}(s) &= \Pr(S_N = s) \\ &= \frac{1}{1 - af_X(0)} \left\{ [p_1 - (a + b)p_0] f_X(s) + \sum_{x=1}^{s \wedge m} \left( a + \frac{bx}{s} \right) f_X(x) f_{S_N}(s - x) \right\}. \end{aligned}$$

If  $N$  is in the  $(a, b, 0)$  class, then  $p_1 = (a + b)p_0$  and so

$$f_{S_N}(s) = \frac{1}{1 - af_X(0)} \left\{ \sum_{x=1}^{s \wedge m} \left( a + \frac{bx}{s} \right) f_X(x) f_{S_N}(s-x) \right\}.$$

**Special Case: Poisson Frequency.** If  $N \sim Poi(\lambda)$ , then  $a = 0$  and  $b = \lambda$ , and thus

$$f_{S_N}(s) = \frac{\lambda}{s} \left\{ \sum_{x=1}^{s \wedge m} x f_X(x) f_{S_N}(s-x) \right\}.$$


---

**Example 7.4.1. Actuarial Exam Question.** The number of claims in a period  $N$  has a geometric distribution with mean 4. The amount of each claim  $X$  follows  $\Pr(X = x) = 0.25$ , for  $x = 1, 2, 3, 4$ . The number of claims and the claim amount are independent.  $S_N$  is the aggregate claim amount in the period. Calculate  $F_{S_N}(3)$ .

**Example Solution.** The severity distribution  $X$  follows

$$f_X(x) = \frac{1}{4}, \quad x = 1, 2, 3, 4.$$

The frequency distribution  $N$  is geometric with mean 4, which is a member of the  $(a, b, 0)$  class with  $b = 0$ ,  $a = \frac{\beta}{1+\beta} = \frac{4}{5}$ , and  $p_0 = \frac{1}{1+\beta} = \frac{1}{5}$ . The support of severity component  $X$  is  $\{1, \dots, m = 4\}$ , discrete and finite. Thus, we can use the recursive method

$$\begin{aligned} f_{S_N}(x) &= 1 \sum_{y=1}^{x \wedge m} (a + 0) f_X(y) f_{S_N}(x-y) \\ &= \frac{4}{5} \sum_{y=1}^{x \wedge m} f_X(y) f_{S_N}(x-y). \end{aligned}$$

Specifically, we have

$$\begin{aligned}
 f_{S_N}(0) &= \Pr(N = 0) = p_0 = \frac{1}{5} \\
 f_{S_N}(1) &= \frac{4}{5} \sum_{y=1}^1 f_X(y)f_{S_N}(1-y) = \frac{4}{5}f_X(1)f_{S_N}(0) \\
 &= \frac{4}{5} \left(\frac{1}{4}\right) \left(\frac{1}{5}\right) = \frac{1}{25} \\
 f_{S_N}(2) &= \frac{4}{5} \sum_{y=1}^2 f_X(y)f_{S_N}(2-y) = \frac{4}{5} [f_X(1)f_{S_N}(1) + f_X(2)f_{S_N}(0)] \\
 &= \frac{4}{5} \left[\frac{1}{4} \left(\frac{1}{25} + \frac{1}{5}\right)\right] = \frac{4}{5} \left(\frac{6}{100}\right) = \frac{6}{125} \\
 f_{S_N}(3) &= \frac{4}{5} [f_X(1)f_{S_N}(2) + f_X(2)f_{S_N}(1) + f_X(3)f_{S_N}(0)] \\
 &= \frac{4}{5} \left[\frac{1}{4} \left(\frac{1}{25} + \frac{1}{5} + \frac{6}{125}\right)\right] = \frac{1}{5} \left(\frac{5+25+6}{125}\right) = 0.0576 \\
 \Rightarrow F_{S_N}(3) &= f_{S_N}(0) + f_{S_N}(1) + f_{S_N}(2) + f_{S_N}(3) = 0.3456.
 \end{aligned}$$

**Example 7.4.2. Convolution Method to Compute the Aggregate Loss Distribution - Continued.** This is a continuation of Example 7.3.4 where we now compute the aggregate loss distribution using the recursive method. This requires discretization of the severity amounts and this illustration rounds claims to the nearest thousand. The following block of code illustrates the calculation.

```

# Discretized severity distribution
round_any = function(y, accuracy, f = round) {
  f(y/accuracy) * accuracy
}
# round to $1000
acc <- 1000
xbar_disc <- round_any(xbar, acc)/acc
dSev <- function(y) {
  re <- ecdf(xbar_disc)(y) - ecdf(xbar_disc)(y - 1)
  re
}

Fs0 <- function(y) {
  if (y < 0)
    return(NA)
  y_scale <- round_any(y, acc)/acc
  y_scale <- ifelse(y_scale > max(xbar_disc), max(xbar_disc), y_scale)
  s.out <- rep(NA, y_scale + 1)
  s.out[1] <- exp(-lambda)
  if (y_scale > 0) {
    for (i in 1:y_scale) {
      s.out[i + 1] <- s.out[i] * dSev(i)
    }
  }
  s.out
}

```

```

    re <- 0
    for (j in 1:i) {
      re <- re + j * dSev(j) * s.out[i + 1 - j]
    }
    s.out[i + 1] <- re * lambda/i
  }
}
return(sum(s.out))
}
Fs <- function(y) sapply(y, Fs0)
obs <- c(-1, 0, 1, 10, 100, 1000, 10000, 100000, 1000000)
# Fs(obs)

```

### 7.4.2 Simulation

The distribution of aggregate loss can be evaluated using Monte Carlo simulation. You can get a broad introduction to simulation procedures in Chapter 8. For aggregate losses, the idea is that one can calculate the empirical distribution of  $S_N$  using a random sample. The expected value and variance of the aggregate loss can also be estimated using the sample mean and sample variance of the simulated values.

We now summarize simulation procedures for aggregate loss models. Let  $R$  be the size of the generated random sample of aggregate losses.

#### 1. Individual Risk Model: $S_n = X_1 + \dots + X_n$

- Let  $j = 1, \dots, R$  be a counter. Start by setting  $j = 1$ .
- Generate each individual loss realization  $x_{ij}$  for  $i = 1, \dots, n$ . For example, this can be done using the inverse transformation method (Section ??).
- Calculate the aggregate loss  $s_j = x_{1j} + \dots + x_{nj}$ .
- Repeat the above two steps for  $j = 2, \dots, R$  to obtain a size- $R$  sample of  $S_n$ , i.e.  $\{s_1, \dots, s_R\}$ .

#### 2. Collective Risk Model: $S_N = X_1 + \dots + X_N$

- Let  $j = 1, \dots, R$  be a counter. Start by setting  $j = 1$ .
- Generate the number of claims  $n_j$  from the frequency distribution  $N$ .
- Given  $n_j$ , generate the amount of each claim independently from severity distribution  $X$ , denoted by  $x_{1j}, \dots, x_{n_j j}$ .
- Calculate the aggregate loss  $s_j = x_{1j} + \dots + x_{n_j j}$ .
- Repeat the above three steps for  $j = 2, \dots, R$  to obtain a size- $R$  sample of  $S_N$ , i.e.  $\{s_1, \dots, s_R\}$ .

Given the random sample of  $S$ , the empirical distribution can be calculated as

$$\hat{F}_S(s) = \frac{1}{R} \sum_{j=1}^R I(s_j \leq s),$$

where  $I(\cdot)$  is an indicator function. The empirical distribution  $\hat{F}_S(s)$  will converge to  $F_S(s)$  almost surely as the sample size  $R \rightarrow \infty$ .

The above procedure assumes that the probability distributions, including the parameter values, of the frequency and severity distributions are known. In practice, one would need to first assume these distributions, estimate their parameters from data, and then assess the quality of model fit using various model validation tools (see Chapter 6). For instance, the assumptions in the collective risk model suggest a two-stage estimation where one model is developed for the number of claims  $N$  from data on claim counts, and another model is developed for the severity of claims  $X$  from data on the amount of claims.

**Example 7.4.3.** Recall Example 7.3.6 with an individual's claim frequency  $N$  has a Poisson distribution with mean  $\lambda = 25$  and claim severity  $X$  is uniformly distributed on the interval  $(5, 95)$ . Using a simulated sample of 10,000 observations, estimate the mean and variance of the aggregate loss  $S_N$ . In addition, use the simulated sample to estimate the probability that aggregate claims for this individual will exceed 2,000 and compare with the normal approximation estimates from Example 7.3.6.

**Solution.** We follow the algorithm for the collective risk model, where we first simulate frequencies  $n_1, \dots, n_{10000}$ , and conditional on  $n_j$ ,  $j = 1, \dots, 10000$ , simulate each individual loss  $x_{ij}$ ,  $i = 1, \dots, n_j$ .

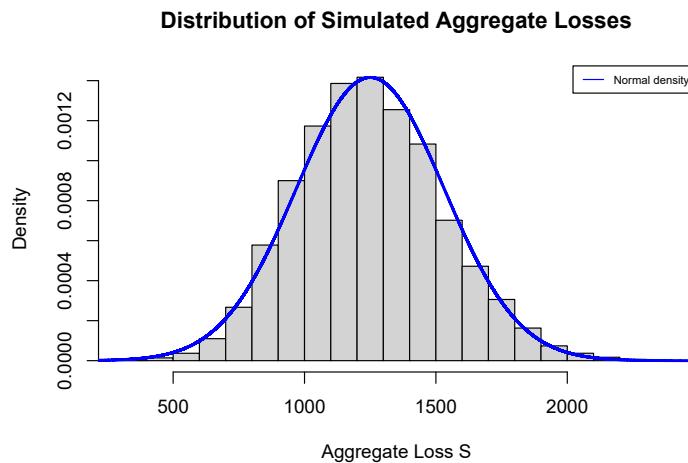
```
set.seed(4321) # For reproducibility of results
m <- 10000 # Number of observations to simulate
lambda <- 25 # Parameter for frequency distribution N
a <- 5
b <- 95 # Parameters for severity distribution X
S <- rep(NA, m) # Initialize an empty vector to store S observations

n <- rpois(m, lambda) # Generate m=10000 observations of N from Poisson
for (j in 1:m) {
    n_j <- n[j] # Given each n_j (j=1,...,m), generate n_j observations of X from uniform
    x_j <- runif(n_j, min = a, max = b)
    s_j <- sum(x_j) # Calculate the aggregate loss s_j
    S[j] <- s_j # Store s_j in the vector of observations
}
# mean(S) # Compare to theoretical value of 1,250 var(S) # Compare to
# theoretical value of 79,375 mean(S>2000) # Proportion of simulated
```

```
# observations s_j that are > 2000 Compare to normal approximation method of
# 0.003884
```

Using simulation, we estimate the mean and variance of the aggregate claims to be approximately 1248 and 77441 respectively, compared to the theoretical values of 1,250 and 79,375. In addition, we estimate the probability that aggregate losses exceed 2000 to be 0.0062, compared to the normal approximation estimate of 0.003884.

We can assess the appropriateness of the normal approximation by comparing the empirical distribution of the simulated aggregate losses to the density of the normal distribution used for the normal approximation,  $N(\mu = 1,250, \sigma^2 = 79,375)$ :



The simulated losses are slightly more right-skewed than the normal distribution, with a longer right tail. This explains why the normal approximation estimate of  $\Pr(S_N > 2000)$  is lower than the simulated estimate.

## 7.5 Effects of Coverage Modifications

In this section, you learn to evaluate:

- the effect of exposure change on the aggregate claim count

- the effect of per-occurrence deductible on the claim frequency
  - the effect of coverage modifications on the aggregate losses
- 

### 7.5.1 Impact of Exposure on Frequency

This section focuses on an individual risk model for claim counts. Recall the individual risk model involves a fixed  $n$  number of contracts and independent loss random variables  $X_i$ . Consider the number of claims from a group of  $n$  policies:

$$S = X_1 + \cdots + X_n,$$

where we assume  $X_i$  are *iid* representing the number of claims from policy  $i$ . In this case, the exposure for the portfolio is  $n$ , using policy as exposure base. In Section ?? we will introduce other exposure bases. The *pgf* of  $S$  is

$$\begin{aligned} P_S(z) &= E(z^S) = E\left(z^{\sum_{i=1}^n X_i}\right) \\ &= \prod_{i=1}^n E(z^{X_i}) = [P_X(z)]^n. \end{aligned}$$

**Special Case: Poisson.** If  $X_i \sim Poi(\lambda)$ , its *pgf* is  $P_X(z) = e^{\lambda(z-1)}$ . Then the *pgf* of  $S$  is

$$P_S(z) = [e^{\lambda(z-1)}]^n = e^{n\lambda(z-1)}.$$

So  $S \sim Poi(n\lambda)$ . That is, the sum of  $n$  independent Poisson random variables each with mean  $\lambda$  has a Poisson distribution with mean  $n\lambda$ .

---

**Special Case: Negative Binomial.** If  $X_i \sim NB(\beta, r)$ , its *pgf* is  $P_X(z) = [1 - \beta(z-1)]^{-r}$ . Then the *pgf* of  $S$  is

$$P_S(z) = [[1 - \beta(z-1)]^{-r}]^n = [1 - \beta(z-1)]^{-nr}.$$

So  $S \sim NB(\beta, nr)$ .

---

**Example 7.5.1.** Assume that the number of claims for each vehicle is Poisson with mean  $\lambda$ . Given the following data on the observed number of claims for each household, calculate the MLE of  $\lambda$ .

Household ID	Number of vehicles	Number of claims
1	2	0
2	1	2
3	3	2
4	1	0
5	1	1

**Example Solution.** Each of the 5 households has number of exposures  $n_j$  (number of vehicles) and number of claims  $S_j$ ,  $j = 1, \dots, 5$ . Note for each household, the number of claims  $S_j \sim Poi(n_j\lambda)$ . The likelihood function is

$$\begin{aligned} L(\lambda) &= \prod_{j=1}^5 \Pr(S_j = s_j) = \prod_{j=1}^5 \frac{e^{-n_j\lambda}(n_j\lambda)^{s_j}}{s_j!} \\ &= \left( \frac{e^{-2\lambda}(2\lambda)^0}{0!} \right) \left( \frac{e^{-1\lambda}(1\lambda)^2}{2!} \right) \left( \frac{e^{-3\lambda}(3\lambda)^2}{2!} \right) \left( \frac{e^{-1\lambda}(1\lambda)^0}{0!} \right) \left( \frac{e^{-1\lambda}(1\lambda)^1}{1!} \right) \\ &\propto e^{-8\lambda}\lambda^5 \end{aligned}$$

Taking the logarithm, we have

$$l(\lambda) = \log L(\lambda) = -8\lambda + 5\log(\lambda).$$

Setting the first derivative of the log-likelihood to 0, we get  $\hat{\lambda} = \frac{5}{8}$ .

If the exposure of the portfolio changes from  $n_1$  to  $n_2$ , we can establish the following relation between the aggregate claim counts:

$$P_{S_{n_2}}(z) = [P_X(z)]^{n_2} = [P_X(z)^{n_1}]^{n_2/n_1} = P_{S_{n_1}}(z)^{n_2/n_1}.$$

### 7.5.2 Impact of Deductibles on Claim Frequency

This section examines the effect of deductibles on claim frequency. Intuitively, there will be fewer claims filed when a policy deductible is imposed because a loss below the deductible level may not result in a claim. Even if an insured does file a claim, this may not result in a payment by the policy, since the claim may be denied or the loss amount may ultimately be determined to be below deductible. Let  $N^L$  denote the number of losses (i.e. the number of claims with no deductible), and  $N^P$  denote the number of payments when a deductible  $d$  is imposed. Our goal is to identify the distribution of  $N^P$  given the distribution of  $N^L$ . We show below that the relationship between  $N^L$  and  $N^P$  can be established within an aggregate risk model framework.

Note that sometimes changes in deductibles will affect policyholder claim behavior. We assume that this is not the case, i.e. the underlying distributions of losses for both frequency and severity remain unchanged when the deductible changes.

Given there are  $N^L$  losses, let  $X_1, X_2 \dots, X_{N^L}$  be the associated amount of losses. For  $j = 1, \dots, N^L$ , define

$$I_j = \begin{cases} 1 & \text{if } X_j > d \\ 0 & \text{otherwise} \end{cases}.$$

Then we establish

$$N^P = I_1 + I_2 + \dots + I_{N^L},$$

that is, the total number of payments is equal to the number of losses above the deductible level. Given that  $I_j$ 's are independent Bernoulli random variables with probability of success  $v = \Pr(X > d)$ , the sum of a *fixed number* of such variables is then a binomial random variable. Thus, conditioning on  $N^L$ ,  $N^P$  has a binomial distribution, i.e.  $N^P | N^L \sim \text{Bin}(N^L, v)$ , where  $v = \Pr(X > d)$ . This implies that

$$\mathbb{E}(z^{N^P} | N^L) = [1 + v(z - 1)]^{N^L}$$

So the *pgf* of  $N^P$  is

$$\begin{aligned} P_{N^P}(z) &= \mathbb{E}_{N^P}(z^{N^P}) = \mathbb{E}_{N^L}[\mathbb{E}_{N^P}(z^{N^P} | N^L)] \\ &= \mathbb{E}_{N^L}[(1 + v(z - 1))^{N^L}] \\ &= P_{N^L}(1 + v(z - 1)). \end{aligned}$$

Thus, we can write the *pgf* of  $N^P$  as the *pgf* of  $N^L$ , evaluated at a new argument  $z^* = 1 + v(z - 1)$ . That is,  $P_{N^P}(z) = P_{N^L}(z^*)$ .

### Special Cases:

- $N^L \sim \text{Poi}(\lambda)$ . The *pgf* of  $N^L$  is  $P_{N^L} = e^{\lambda(z-1)}$ . Thus the *pgf* of  $N^P$  is

$$\begin{aligned} P_{N^P}(z) &= e^{\lambda(1+v(z-1)-1)} \\ &= e^{\lambda v(z-1)}, \end{aligned}$$

So  $N^P \sim \text{Poi}(\lambda v)$ . This means the number of payments has the same distribution as the number of losses, but with the expected number of payments equal to  $\lambda v = \lambda \Pr(X > d)$ .

- $N^L \sim NB(\beta, r)$ . The pgf of  $N^L$  is  $P_{N^L}(z) = [1 - \beta(z - 1)]^{-r}$ . Thus the pgf of  $N^P$  is

$$\begin{aligned} P_{N^P}(z) &= (1 - \beta(1 + v(z - 1) - 1))^{-r} \\ &= (1 - \beta v(z - 1))^{-r}, \end{aligned}$$

So  $N^P \sim NB(\beta v, r)$ . This means the number of payments has the same distribution as the number of losses, but with parameters  $\beta v$  and  $r$ .

**Example 7.5.2.** Suppose that loss amounts  $X_i \sim \text{Pareto}(\alpha = 4, \theta = 150)$ . You are given that the loss frequency is  $N^L \sim \text{Poi}(\lambda)$  and the payment frequency distribution is  $N_1^P \sim \text{Poi}(0.4)$  at deductible level  $d_1 = 30$ . Find the distribution of the payment frequency  $N_2^P$  when the deductible level is  $d_2 = 100$ .

**Example Solution.** Because the loss frequency  $N^L$  is Poisson, we can relate the means of the loss distribution  $N^L$  and the first payment distribution  $N_1^P$  (under deductible  $d_1 = 30$ ) through  $0.4 = \lambda v_1$ , where

$$\begin{aligned} v_1 &= \Pr(X > 30) = \left( \frac{150}{30 + 150} \right)^4 = \left( \frac{5}{6} \right)^4 \\ \Rightarrow \lambda &= 0.4 \left( \frac{6}{5} \right)^4. \end{aligned}$$

With this, we can assess the second payment distribution  $N_2^P$  (under deductible  $d_2 = 100$ ) as being Poisson with mean  $\lambda_2 = \lambda v_2$ , where

$$\begin{aligned} v_2 &= \Pr(X > 100) = \left( \frac{150}{100 + 150} \right)^4 = \left( \frac{3}{5} \right)^4 \\ \Rightarrow \lambda_2 &= \lambda v_2 = 0.4 \left( \frac{6}{5} \right)^4 \left( \frac{3}{5} \right)^4 = 0.1075. \end{aligned}$$

**Example 7.5.3. Follow-Up.** Now suppose instead that the loss frequency is  $N^L \sim NB(\beta, r)$  and for deductible  $d_1 = 30$ , the payment frequency  $N_1^P$  is negative binomial with mean 0.4. Find the mean of the payment frequency  $N_2^P$  for deductible  $d_2 = 100$ .

**Example Solution.** Because the loss frequency  $N^L$  is negative binomial, we can relate the parameter  $\beta$  of the  $N^L$  distribution and the parameter  $\beta_1$  of the

first payment distribution  $N_1^P$  using  $\beta_1 = \beta v_1$ , where

$$v_1 = \Pr(X > 30) = \left(\frac{5}{6}\right)^4$$

Thus, the mean of  $N_1^P$  and the mean of  $N^L$  are related via

$$\begin{aligned} 0.4 &= r\beta_1 = r(\beta v_1) \\ \Rightarrow r\beta &= \frac{0.4}{v_1} = 0.4 \left(\frac{6}{5}\right)^4. \end{aligned}$$

Note that  $v_2 = \Pr(X > 100) = \left(\frac{3}{5}\right)^4$  as in the original example. Then the second payment frequency distribution under deductible  $d_2 = 100$  is  $N_2^P \sim NB(\beta v_2, r)$  with mean

$$r(\beta v_2) = (r\beta)v_2 = 0.4 \left(\frac{6}{5}\right)^4 \left(\frac{3}{5}\right)^4 = 0.1075.$$

Next, we examine the more general case where  $N^L$  is a zero-modified distribution. Recall that a zero-modified distribution can be defined in terms of an unmodified one (as was shown in Section 3.5.1). That is,

$$p_k^M = c p_k^0, \text{ for } k = 1, 2, 3, \dots, \text{ with } c = \frac{1 - p_0^M}{1 - p_0^0},$$

where  $p_k^0$  is the pmf of the unmodified distribution. In the case that  $p_0^M = 0$ , we call this a *zero-truncated* distribution, or *ZT*. For other arbitrary values of  $p_0^M$ , this is a zero-modified, or *ZM*, distribution. The pgf for the modified distribution is shown as

$$P^M(z) = 1 - c + c P^0(z),$$

expressed in terms of the pgf of the unmodified distribution,  $P^0(z)$ . When  $N^L$  follows a zero-modified distribution, the distribution of  $N^P$  is established using the same relation from earlier,  $P_{NP}(z) = P_{NL}(1 + v(z-1))$ .

### Special Cases:

- $N^L$  is a ZM-Poisson random variable with parameters  $\lambda$  and  $p_0^M$ . The pgf of  $N^L$  is

$$P_{NL}(z) = 1 - \frac{1 - p_0^M}{1 - e^{-\lambda}} + \frac{1 - p_0^M}{1 - e^{-\lambda}} (e^{\lambda(z-1)}).$$

Thus the pgf of  $N^P$  is

$$P_{NP}(z) = 1 - \frac{1 - p_0^M}{1 - e^{-\lambda}} + \frac{1 - p_0^M}{1 - e^{-\lambda}} (e^{\lambda v(z-1)}).$$

So the number of payments is also a ZM-Poisson distribution with parameters  $\lambda v$  and  $p_0^M$ . The probability at zero can be evaluated using  $\Pr(N^P = 0) = P_{NP}(0)$ .

- $N^L$  is a ZM-negative binomial random variable with parameters  $\beta$ ,  $r$ , and  $p_0^M$ . The pgf of  $N^L$  is

$$P_{NL}(z) = 1 - \frac{1 - p_0^M}{1 - (1 + \beta)^{-r}} + \frac{1 - p_0^M}{1 - (1 + \beta)^{-r}} [1 - \beta(z - 1)]^{-r}.$$

Thus the pgf of  $N^P$  is

$$P_{NP}(z) = 1 - \frac{1 - p_0^M}{1 - (1 + \beta)^{-r}} + \frac{1 - p_0^M}{1 - (1 + \beta)^{-r}} [1 - \beta v(z - 1)]^{-r}.$$

So the number of payments is also a ZM-negative binomial distribution with parameters  $\beta v$ ,  $r$ , and  $p_0^M$ . Similarly, the probability at zero can be evaluated using  $\Pr(N^P = 0) = P_{NP}(0)$ .

**Example 7.5.4.** Aggregate losses are modeled as follows:

- (i) The number of losses follows a zero-modified Poisson distribution with  $\lambda = 3$  and  $p_0^M = 0.5$ .
- (ii) The amount of each loss has a Burr distribution with  $\alpha = 3, \theta = 50, \gamma = 1$ .
- (iii) There is a deductible of  $d = 30$  on each loss.
- (iv) The number of losses and the amounts of the losses are mutually independent.

Calculate  $E(N^P)$  and  $\text{Var}(N^P)$ .

**Example Solution.** Since  $N^L$  follows a ZM-Poisson distribution with parameters  $\lambda$  and  $p_0^M$ , we know that  $N^P$  also follows a ZM-Poisson distribution, but with parameters  $\lambda v$  and  $p_0^M$ , where

$$v = \Pr(X > 30) = \left( \frac{1}{1 + (30/50)} \right)^3 = 0.2441.$$

Thus,  $N^P$  follows a ZM-Poisson distribution with parameters  $\lambda^* = \lambda v = 0.7324$

and  $p_0^M = 0.5$ . Finally,

$$\begin{aligned}\mathbb{E}(N^P) &= (1 - p_0^M) \frac{\lambda^*}{1 - e^{-\lambda^*}} = 0.5 \left( \frac{0.7324}{1 - e^{-0.7324}} \right) \\ &= 0.7053 \\ \text{Var}(N^P) &= (1 - p_0^M) \left( \frac{\lambda^*[1 - (\lambda^* + 1)e^{-\lambda^*}]}{(1 - e^{-\lambda^*})^2} \right) + p_0^M(1 - p_0^M) \left( \frac{\lambda^*}{1 - e^{-\lambda^*}} \right)^2 \\ &= 0.5 \left( \frac{0.7324(1 - 1.7324e^{-0.7324})}{(1 - e^{-0.7324})^2} \right) + 0.5^2 \left( \frac{0.7324}{1 - e^{-0.7324}} \right)^2 \\ &= 0.7244.\end{aligned}$$

### 7.5.3 Impact of Policy Modifications on Aggregate Claims

In this section, we examine how a change in the deductible affects the aggregate payments from an insurance portfolio. We assume that the presence of policy limits ( $u$ ), coinsurance ( $\alpha$ ), and inflation ( $r$ ) have no effect on the underlying distribution of frequency of payments made by an insurer. As in the previous section, we further assume that deductible changes do not impact the underlying distributions of losses for both frequency and severity.

Recall the notation  $N^L$  for the number of losses. With ground-up loss amount  $X$  and policy deductible  $d$ , we use  $N^P$  for the number of payments (as defined in the previous section 7.5.2). Also, define the amount of payment on a per-loss basis as

$$Y^L = \begin{cases} 0 , & \text{if } X < \frac{d}{1+r} \\ \alpha[(1+r)X - d] , & \text{if } \frac{d}{1+r} \leq X < \frac{u}{1+r} , \\ \alpha(u-d) , & \text{if } X \geq \frac{u}{1+r} \end{cases}$$

and the amount of payment on a per-payment basis as

$$Y^P = \begin{cases} \text{undefined} , & \text{if } X < \frac{d}{1+r} \\ \alpha[(1+r)X - d] , & \text{if } \frac{d}{1+r} \leq X < \frac{u}{1+r} . \\ \alpha(u-d) , & \text{if } X \geq \frac{u}{1+r}. \end{cases}$$

In the above,  $r$ ,  $u$ , and  $\alpha$  represent the inflation rate, policy limit, and coinsurance, respectively. Hence, aggregate costs (payment amounts) can be ex-

pressed either on a per loss or per payment basis:

$$\begin{aligned} S &= Y_1^L + \cdots + Y_{N_L}^L \\ &= Y_1^P + \cdots + Y_{N_P}^P. \end{aligned}$$

(Recall that when we introduced the per-loss and per-payment bases in Section 5.1, we used another letter  $Y$  to distinguish losses from insurance payments, or claims. At this point in our development, we use the letter  $X$  to reduce notation complexity.)

The fundamentals regarding collective risk models are ready to apply. For instance, we have:

$$\begin{aligned} E(S) &= E(N^L) E(X^L) = E(N^P) E(Y^P) \\ \text{Var}(S) &= E(N^L) \text{Var}(Y^L) + [E(Y^L)]^2 \text{Var}(N^L) \\ &= E(N^P) \text{Var}(Y^P) + [E(Y^P)]^2 \text{Var}(N^P) \\ M_S(z) &= P_{N^L} [M_{Y^L}(z)] = P_{N^P} [M_{Y^P}(z)]. \end{aligned}$$

**Example 7.5.5. Actuarial Exam Question.** A group dental policy has a negative binomial claim count distribution with mean 300 and variance 800. Ground-up severity is given by the following table:

Severity	Probability
40	0.25
80	0.25
120	0.25
200	0.25

You expect severity to increase 50% with no change in frequency. You decide to impose a per claim deductible of 100. Calculate the expected total claim payment  $S$  after these changes.

**Example Solution.** The cost per loss with a 50% increase in severity and a 100 deductible per claim is

$$X^L = \begin{cases} 0 & 1.5x < 100 \\ 1.5x - 100 & 1.5x \geq 100 \end{cases}$$

This has expectation

$$\begin{aligned} E(X^L) &= \frac{1}{4} [(1.5(40) - 100)_+ + (1.5(80) - 100)_+ + (1.5(120) - 100)_+ + (1.5(200) - 100)_+] \\ &= \frac{1}{4} [(60 - 100)_+ + (120 - 100)_+ + (180 - 100)_+ + (300 - 100)_+] \\ &= \frac{1}{4} [0 + 20 + 80 + 200] = 75. \end{aligned}$$

Thus, the expected aggregate loss is

$$E(S) = E(N) E(X^L) = 300(75) = 22,500..$$

**Example 7.5.6. Follow-Up.** What is the variance of the total claim payment,  $\text{Var}(S)$ ?

**Example Solution.** On a per loss basis, we have

$$\text{Var}(S) = E(N) \text{Var}(X^L) + [E(X^L)]^2 \text{Var}(N)$$

where  $E(N) = 300$  and  $\text{Var}(N) = 800$ . We find

$$\begin{aligned} E[(X^L)^2] &= \frac{1}{4} [0^2 + 20^2 + 80^2 + 200^2] = 11,700 \\ \Rightarrow \text{Var}(X^L) &= E[(X^L)^2] - [E(X^L)]^2 = 11,700 - 75^2 = 6,075 \end{aligned}$$

Thus, the variance of the aggregate claim payment is

$$\text{Var}(S) = 300(6,075) + 75^2(800) = 6,322,500.$$

*Alternative Method: Using the Per Payment Basis.* Previously, we calculated the expected total claim payment by multiplying the expected number of losses by the expected payment *per loss*. Recall that we can also multiply the expected number of payments by the expected payment *per payment*. In this case, we have

$$S = X_1^P + \cdots + X_{N_P}^P$$

The probability of a payment is

$$\Pr(1.5X \geq 100) = \Pr(X \geq 66.\bar{6}) = \frac{3}{4}.$$

Thus, the number of payments,  $N^P$  has a negative binomial distribution (see

negative binomial special case in Section 7.5.2) with mean

$$\mathbb{E}(N^P) = \mathbb{E}(N^L) \Pr(1.5X \geq 100) = 300 \left(\frac{3}{4}\right) = 225.$$

The cost per payment is

$$X^P = \begin{cases} \text{undefined , if } 1.5x < 100 \\ 1.5x - 100 , \text{ if } 1.5x \geq 100 \end{cases}$$

This has expectation

$$\mathbb{E}(X^P) = \frac{\mathbb{E}(X^L)}{\Pr(1.5X > 100)} = \frac{75}{(3/4)} = 100.$$

Thus, as before, the expected aggregate loss is

$$\mathbb{E}(S) = \mathbb{E}(X^P) \mathbb{E}(N^P) = 100(225) = 22,500.$$

**Example 7.5.7. Actuarial Exam Question.** A company insures a fleet of vehicles. Aggregate losses have a compound Poisson distribution. The expected number of losses is 20. Loss amounts, regardless of vehicle type, have exponential distribution with  $\theta = 200$ . To reduce the cost of the insurance, two modifications are to be made:

- (i) A certain type of vehicle will not be insured. It is estimated that this will reduce loss frequency by 20%.
- (ii) A deductible of 100 per loss will be imposed.

Calculate the expected aggregate amount paid by the insurer after the modifications.

**Example Solution.** On a per loss basis, we have a 100 deductible. Thus, the expectation per loss is

$$\begin{aligned} \mathbb{E}(X^L) &= E[(X - 100)_+] = E(X) - E(X \wedge 100) \\ &= 200 - 200(1 - e^{-100/200}) = 121.31. \end{aligned}$$

Loss frequency has been reduced by 20%, resulting in an expected number of losses

$$\mathbb{E}(N^L) = 0.8(20) = 16.$$

Thus, the expected aggregate amount paid after the modifications is

$$\mathbb{E}(S) = \mathbb{E}(X^L) \mathbb{E}(N^L) = 121.31(16) = 1,941.$$

*Alternative Method: Using the Per Payment Basis.* We can also use the per payment basis to find the expected aggregate amount paid after the modifications. With the deductible of 100, the probability that a payment occurs is  $\Pr(X > 100) = e^{-100/200}$ . For the per payment severity, plugging in the expression for  $E(X^L)$  from the original example, we have

$$E(X^P) = \frac{E(X^L)}{\Pr(X > 100)} = \frac{200 - 200(1 - e^{-100/200})}{e^{-100/200}} = 200$$

This is not surprising – recall that the exponential distribution is memoryless, so the expected claim amount paid in excess of 100 is still exponential with mean 200.

Now we look at the payment frequency

$$E(N^P) = E(N^L) \Pr(X > 100) = 16 e^{-100/200} = 9.7.$$

Putting this together, we produce the same answer using the per payment basis as the per loss basis from earlier

$$E(S) = E(X^P) E(N^P) = 200(9.7) = 1,941.$$


---



---

## 7.6 Further Resources and Contributors

### Contributors

- **Peng Shi** and **Lisa Gao**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter.
- **Peng Shi**, University of Wisconsin-Madison, is the author of the second edition of this chapter. Email: [pshi@bus.wisc.edu](mailto:pshi@bus.wisc.edu) for chapter comments and suggested improvements.
- Chapter reviewers include: Himchan Jeong, Vytaras Brazauskas, Mark Maxwell, Jiadong Ren, Sherly Paola Alfonso Sanchez, and Di (Cindy) Xu.

### Further Readings and References

If you would like additional practice with R coding, please visit our companion [LDA Short Course](#). In particular, see the [Aggregate Loss Models Chapter](#).

### TS 7.A.1. Individual Risk Model Properties

For the expected value of the aggregate loss under the individual risk model,

$$\begin{aligned} E(S_n) &= \sum_{i=1}^n E(X_i) = \sum_{i=1}^n E(I_i \times B_i) = \sum_{i=1}^n E(I_i) E(B_i) \quad \text{from independence of } I_i \text{'s and } B_i \text{'s} \\ &= \sum_{i=1}^n \Pr(I_i = 1) \mu_i \quad \text{since } E(I_i) \text{ is the probability } I_i \text{ is 1} \\ &= \sum_{i=1}^n q_i \mu_i. \end{aligned}$$

For the variance of the aggregate loss under the individual risk model,

$$\begin{aligned} \text{Var}(S_n) &= \sum_{i=1}^n \text{Var}(X_i) \quad \text{from independence of } X_i \text{'s} \\ &= \sum_{i=1}^n (E[\text{Var}(X_i|I_i)] + \text{Var}[E(X_i|I_i)]) \quad \text{from conditional variance formulas} \\ &= \sum_{i=1}^n (q_i \sigma_i^2 + q_i (1 - q_i) \mu_i^2). \end{aligned}$$

To see this, note that

$$\begin{aligned} E[\text{Var}(X_i|I_i)] &= \text{Var}(X_i|I_i = 0) \Pr(I_i = 0) + \text{Var}(X_i|I_i = 1) \Pr(I_i = 1) \\ &= q_i \sigma_i^2 + (1 - q_i) (0) = q_i \sigma_i^2, \end{aligned}$$

and

$$\text{Var}[E(X_i|I_i)] = q_i (1 - q_i) \mu_i^2,$$

using the Bernoulli variance shortcut since  $E(X_i|I_i) = 0$  when  $I_i = 0$  (probability  $\Pr(I_i = 0) = 1 - q_i$ ) and  $E(X_i|I_i) = \mu_i$  when  $I_i = 1$  (probability  $\Pr(I_i = 1) = q_i$ ).

For the probability generating function of the aggregate loss under the individual risk model,

$$\begin{aligned} P_{S_n}(z) &= \prod_{i=1}^n P_{X_i}(z) \quad \text{from the independence of } X_i \text{'s} \\ &= \prod_{i=1}^n E(z^{X_i}) = \prod_{i=1}^n E(z^{I_i \times B_i}) = \prod_{i=1}^n E[E(z^{I_i \times B_i}|I_i)] \quad \text{from law of iterated expectations} \\ &= \prod_{i=1}^n [E(z^{I_i \times B_i}|I_i = 0) \Pr(I_i = 0) + E(z^{I_i \times B_i}|I_i = 1) \Pr(I_i = 1)] \\ &= \prod_{i=1}^n [(1)(1 - q_i) + P_{B_i}(z) q_i] = \prod_{i=1}^n (1 - q_i + q_i P_{B_i}(z)) \end{aligned}$$

Lastly, for the moment generating function of the aggregate loss under the individual risk model,

$$\begin{aligned}
 M_{S_n}(t) &= \prod_{i=1}^n M_{X_i}(t) \quad \text{from the independence of } X_i\text{'s} \\
 &= \prod_{i=1}^n E(e^{t X_i}) = \prod_{i=1}^n E(e^{t (I_i \times B_i)}) \\
 &= \prod_{i=1}^n E [E(e^{t (I_i \times B_i)} | I_i)] \quad \text{from law of iterated expectations} \\
 &= \prod_{i=1}^n [E(e^{t (I_i \times B_i)} | I_i = 0) \Pr(I_i = 0) + E(e^{t (I_i \times B_i)} | I_i = 1) \Pr(I_i = 1)] \\
 &= \prod_{i=1}^n [(1)(1 - q_i) + M_{B_i}(t) q_i] = \prod_{i=1}^n (1 - q_i + q_i M_{B_i}(t)).
 \end{aligned}$$


---

#### TS 7.A.2. Relationship Between Probability Generating Functions of $X_i$ and $X_i^T$

Let  $X_i$  belong to the  $(a, b, 0)$  class with pmf  $p_{ik} = \Pr(X_i = k)$  for  $k = 0, 1, \dots$  and  $X_i^T$  be the associated zero-truncated distribution in the  $(a, b, 1)$  class with pmf  $p_{ik}^T = p_{ik}/(1 - p_{i0})$  for  $k = 1, 2, \dots$ . Then the relationship between the pgf of  $X_i$  and the pgf of  $X_i^T$  is shown by

$$\begin{aligned}
 P_{X_i}(z) &= E(z^{X_i}) = E[E(z^{X_i} | X_i)] \quad \text{from law of iterated expectations} \\
 &= E(z^{X_i} | X_i = 0) \Pr(X_i = 0) + E(z^{X_i} | X_i > 0) \Pr(X_i > 0) \\
 &= (1) p_{i0} + E(z^{X_i^T}) (1 - p_{i0}) \quad \text{since } (X_i | X_i > 0) \text{ is zero-truncated r.v. } X_i^T \\
 &= p_{i0} + (1 - p_{i0}) P_{X_i^T}(z).
 \end{aligned}$$


---

#### TS 7.A.3. Moment Generating Function of Aggregate Loss $S_N$ in Example 7.3.9

For  $N \sim Geo(\beta)$  and  $X \sim Exp(\theta)$ , we have

$$\begin{aligned}
 P_N(z) &= \frac{1}{1 - \beta(z - 1)} \\
 M_X(t) &= \frac{1}{1 - \theta t}.
 \end{aligned}$$

Thus, the *mgf* of aggregate loss  $S_N$  is

$$\begin{aligned}
 M_{S_N}(t) &= P_N[M_X(t)] = \frac{1}{1 - \beta \left( \frac{1}{1-\theta t} - 1 \right)} \\
 &= \frac{1}{1 - \beta \left( \frac{\theta t}{1-\theta t} \right)} + 1 - 1 = 1 + \frac{\beta \left( \frac{\theta t}{1-\theta t} \right)}{1 - \beta \left( \frac{\theta t}{1-\theta t} \right)} \\
 &= 1 + \frac{\beta \theta t}{(1 - \theta t) - \beta \theta t} = 1 + \frac{\beta \theta t}{1 - \theta t(1 + \beta)} \cdot \frac{1 + \beta}{1 + \beta} \\
 &= 1 + \frac{\beta}{1 + \beta} \left[ \frac{\theta(1 + \beta)t}{1 - \theta(1 + \beta)t} \right] \\
 &= 1 + \frac{\beta}{1 + \beta} \left[ \frac{1}{1 - \theta(1 + \beta)t} - 1 \right],
 \end{aligned}$$

which gives the expression (7.1). For the alternate expression of the *mgf* (7.2), we continue from where we just left off:

$$\begin{aligned}
 M_{S_N}(t) &= 1 + \frac{\beta}{1 + \beta} \left[ \frac{\theta(1 + \beta)t}{1 - \theta(1 + \beta)t} \right] \\
 &= \frac{1 + \beta}{1 + \beta} + \frac{\beta}{1 + \beta} \left[ \frac{\theta(1 + \beta)t}{1 - \theta(1 + \beta)t} \right] \\
 &= \frac{1}{1 + \beta} + \frac{\beta}{1 + \beta} + \frac{\beta}{1 + \beta} \left[ \frac{\theta(1 + \beta)t}{1 - \theta(1 + \beta)t} \right] \\
 &= \frac{1}{1 + \beta} + \frac{\beta}{1 + \beta} \left[ 1 + \frac{\theta(1 + \beta)t}{1 - \theta(1 + \beta)t} \right] \\
 &= \frac{1}{1 + \beta} + \frac{\beta}{1 + \beta} \left[ \frac{1}{1 - \theta(1 + \beta)t} \right].
 \end{aligned}$$



# 8

---

## *Simulation and Resampling*

---

*Chapter Preview.* Simulation is a computationally intensive method used to solve difficult problems. Instead of creating physical processes and experimenting with them in order to understand their operational characteristics, a simulation study is based on a computer representation - it considers various hypothetical conditions as inputs and summarizes the results. Through simulation, a vast number of hypothetical conditions can be quickly and inexpensively examined. Section 8.1 introduces simulation as a valuable computational tool, particularly effective in complex, multivariate settings.

Analysts find simulation especially useful for computing measures that summarize intricate distributions, as discussed in Section 8.2. This encompasses all the examples mentioned in the book thus far, such as measures that summarize the frequency and severity of losses, along with many additional cases. Simulation can also be used to compute complex distributions necessary for hypothesis testing. In addition, we can also use simulation to draw from an empirical distribution - this process is known as *resampling*. Resampling allows us to assess the uncertainty of estimates in complex models. Section 8.3 introduces resampling in the context of bootstrapping to determine the precision of estimators. Section 8.4 on cross-validation shows how to use it for model selection and validation.

---

### 8.1 Random Number Generation

---

In this section, you learn how to:

- Generate approximately independent realizations that are uniformly distributed
- Transform the uniformly distributed realizations to observations from a probability distribution of interest
- Generate simulated values directly from common distributions using ready-made random number generators

- Generate simulated values from complex distributions by combining simulated values from common distributions
  - Generate simulated values from distributions whose domain is restricted to specific regions of interest, such as with deductible and long-tailed actuarial applications.
- 

### 8.1.1 Generating Independent Uniform Observations

The simulations that we consider are generated by computers. A major strength of this approach is that they can be replicated, allowing us to check and improve our work. Naturally, this also means that they are not really random. Nonetheless, algorithms have been produced so that results appear to be random for all practical purposes. Specifically, they pass sophisticated tests of independence and can be designed so that they come from a single distribution - our iid assumption, identically and independently distributed.

To get a sense as to what these algorithms do, we consider a historically prominent method.

**Linear Congruential Generator.** To generate a sequence of random numbers, start with  $B_0$ , a starting value that is known as a *seed*. This value is updated using the recursive relationship

$$B_{n+1} = (aB_n + c) \text{ modulo } m, \quad n = 0, 1, 2, \dots$$

This algorithm is called a linear congruential generator. The case of  $c = 0$  is called a *multiplicative* congruential generator; it is particularly useful for really fast computations.

For illustrative values of  $a$  and  $m$ , Microsoft's Visual Basic uses  $m = 2^{24}$ ,  $a = 1,140,671,485$ , and  $c = 12,820,163$  (see [https://en.wikipedia.org/wiki/Linear\\_congruential\\_generator](https://en.wikipedia.org/wiki/Linear_congruential_generator)). This is the engine underlying the random number generation in Microsoft's Excel program.

The sequence used by the analyst is defined as  $U_n = B_n/m$ . The analyst may interpret the sequence  $\{U_i\}$  to be (approximately) identically and independently uniformly distributed on the interval  $(0,1)$ . To illustrate the algorithm, consider the following.

**Example 8.1.1. Illustrative Sequence.** Take  $m = 15$ ,  $a = 3$ ,  $c = 2$  and  $B_0 = 1$ . Then we have:

step	$n$	$B_n$	$U_n$
0		$B_0 = 1$	
1		$B_1 = \text{mod}(3 \times 1 + 2) = 5$	$U_1 = \frac{5}{15}$
2		$B_2 = \text{mod}(3 \times 5 + 2) = 2$	$U_2 = \frac{2}{15}$
3		$B_3 = \text{mod}(3 \times 2 + 2) = 8$	$U_3 = \frac{8}{15}$
4		$B_4 = \text{mod}(3 \times 8 + 2) = 11$	$U_4 = \frac{11}{15}$

The linear congruential generator is just one method of producing pseudo-random outcomes. It is easy to understand and is widely used. The linear congruential generator does have limitations, including the fact that it is possible to detect long-run patterns over time in the sequences generated (recall that we can interpret *independence* to mean a total lack of functional patterns). Not surprisingly, advanced techniques have been developed that address some of this method's drawbacks. The random number generated by R utilizes such advanced techniques.

Sometimes computer generated random results are known as pseudo-random numbers to reflect the fact that they are machine generated and can be replicated. That is, despite the fact that  $\{U_i\}$  appears to be i.i.d, it can be reproduced by using the same seed number (and the same algorithm).

**Example 8.1.2. Generating Uniform Random Numbers in R.** The following code shows how to generate three uniform (0,1) numbers in R using the `runif` command. The `set.seed()` function sets the initial seed. In many computer packages, the initial seed is set using the system clock unless specified otherwise.

#### Three Uniform Random Variates

```
set.seed(2017)
U <- runif(3)
knitr::kable(U, digits = 5, align = "c", col.names = "Uniform") %>%
  kableExtra::kable_classic(full_width = F) %>%
  kable_styling(latex_options = "hold_position", font_size = 10)
```

Uniform
0.92424
0.53718
0.46920

### 8.1.2 Inverse Transform Method

With the sequence of uniform random numbers, we next transform them to a distribution of interest, say  $F$ . A prominent technique is the inverse transform method, defined as

$$X_i = F^{-1}(U_i).$$

Here, recall from Section 4.1.1 that we introduced the inverse of the distribution function,  $F^{-1}$ , and referred to it also as the quantile function. Specifically, it is defined to be

$$F^{-1}(y) = \inf_x \{F(x) \geq y\}.$$

Recall that  $\inf$  stands for *infimum* or the greatest lower bound. It is essentially the smallest value of  $x$  that satisfies the inequality  $\{F(x) \geq y\}$ . The result is that the sequence  $\{X_i\}$  is *iid* with distribution function  $F$  if the  $\{U_i\}$  are *iid* with uniform on  $(0, 1)$  distribution function.

The inverse transform result is available when the underlying random variable is continuous, discrete or a hybrid combination of the two. We now present a series of examples to illustrate its scope of applications.

**Example 8.1.3. Generating Exponential Random Numbers.** Suppose that we would like to generate observations from an exponential distribution with scale parameter  $\theta$  so that  $F(x) = 1 - e^{-x/\theta}$ . To compute the inverse transform, we can use the following steps:

$$\begin{aligned} y = F(x) &\Leftrightarrow y = 1 - e^{-x/\theta} \\ &\Leftrightarrow -\theta \ln(1 - y) = x = F^{-1}(y). \end{aligned}$$

Thus, if  $U$  has a uniform  $(0, 1)$  distribution, then  $X = -\theta \ln(1 - U)$  has an exponential distribution with parameter  $\theta$ .

The following R code shows how we can start with the same three uniform random numbers as in [Example 8.1.2](#) and transform them to independent exponentially distributed random variables with a mean of 10. Alternatively, you can directly use the `rexp` function in R to generate random numbers from the exponential distribution. The algorithm built into this routine is different so even with the same starting seed number, individual realizations will differ.

```
set.seed(2017)
U <- runif(3)
X1 <- -10 * log(1 - U)
set.seed(2017)
X2 <- rexp(3, rate = 1/10)
```

Uniform	Exponential 1	Exponential 2
0.92424	25.80219	3.25222
0.53718	7.70409	8.47652
0.46920	6.33362	5.40176

Three Uniform Random Variates

---

**Example 8.1.4. Generating Pareto Random Numbers.** Suppose that we would like to generate observations from a Pareto distribution with parameters  $\alpha$  and  $\theta$  so that  $F(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^\alpha$ . To compute the inverse transform, we can use the following steps:

$$\begin{aligned} y = F(x) &\Leftrightarrow 1 - y = \left(\frac{\theta}{x + \theta}\right)^\alpha \\ &\Leftrightarrow (1 - y)^{-1/\alpha} = \frac{x + \theta}{\theta} = \frac{x}{\theta} + 1 \\ &\Leftrightarrow \theta((1 - y)^{-1/\alpha} - 1) = x = F^{-1}(y). \end{aligned}$$

Thus,  $X = \theta((1 - U)^{-1/\alpha} - 1)$  has a Pareto distribution with parameters  $\alpha$  and  $\theta$ .

---

**Inverse Transform Justification.** Why does the random variable  $X = F^{-1}(U)$  have a distribution function  $F$ ?

This is easy to establish when  $F$  is strictly increasing, where the distribution is continuous. Because  $U$  is a uniform random variable on  $(0,1)$ , we know that  $\Pr(U \leq y) = y$ , for  $0 \leq y \leq 1$ . Thus,

$$\begin{aligned} \Pr[X \leq x] &= \Pr[F^{-1}(U) \leq x] \\ &= \Pr[F(F^{-1}(U)) \leq F(x)] \\ &= \Pr[U \leq F(x)] = F(x), \end{aligned}$$

as required. The key step is that  $F[F^{-1}(u)] = u$  for each  $u$ , which is true when  $F$  is strictly increasing.

We now consider some discrete examples.

**Example 8.1.5. Generating Bernoulli Random Numbers.** Suppose that we wish to simulate random variables from a Bernoulli distribution with parameter  $q = 0.85$ .

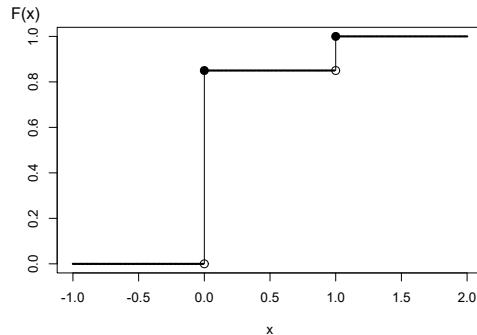


FIGURE 8.1: Distribution Function of a Binary Random Variable

A graph of the cumulative distribution function in Figure 8.1 shows that the quantile function can be written as

$$F^{-1}(y) = \begin{cases} 0 & 0 < y \leq 0.85 \\ 1 & 0.85 < y \leq 1.0. \end{cases}$$

Thus, with the inverse transform we may define

$$X = \begin{cases} 0 & 0 < U \leq 0.85 \\ 1 & 0.85 < U \leq 1.0 \end{cases}$$

For illustration, we generate three random numbers to get

```
set.seed(2017)
U <- runif(3)
X <- 1 * (U > 0.85)
```

*Three Random Variates*

Uniform	Binary X
0.92424	1
0.53718	0
0.46920	0

**Example 8.1.6. Generating Random Numbers from a Discrete Distribution.** Consider the time of a machine failure in the first five years. The distribution of failure times is given as:

Time	1.0	2.0	3.0	4.0	5.0
Probability	0.1	0.2	0.1	0.4	0.2
Distribution Function	0.1	0.3	0.4	0.8	1.0

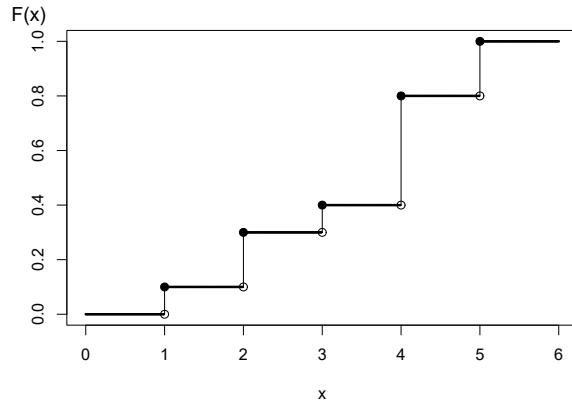


FIGURE 8.2: Distribution Function of a Discrete Random Variable

*Discrete Distribution*

Using the graph of the distribution function in Figure 8.2, with the inverse transform we may define

$$X = \begin{cases} 1 & 0 < U \leq 0.1 \\ 2 & 0.1 < U \leq 0.3 \\ 3 & 0.3 < U \leq 0.4 \\ 4 & 0.4 < U \leq 0.8 \\ 5 & 0.8 < U \leq 1.0. \end{cases}$$

For general discrete random variables there may not be an ordering of outcomes. For example, a person could own one of five types of life insurance products and we might use the following algorithm to generate random outcomes:

$$X = \begin{cases} \text{whole life} & 0 < U \leq 0.1 \\ \text{endowment} & 0.1 < U \leq 0.3 \\ \text{term life} & 0.3 < U \leq 0.4 \\ \text{universal life} & 0.4 < U \leq 0.8 \\ \text{variable life} & 0.8 < U \leq 1.0. \end{cases}$$

Another analyst may use an alternative procedure such as:

$$X = \begin{cases} \text{whole life} & 0.9 < U < 1.0 \\ \text{endowment} & 0.7 \leq U < 0.9 \\ \text{term life} & 0.6 \leq U < 0.7 \\ \text{universal life} & 0.2 \leq U < 0.6 \\ \text{variable life} & 0 \leq U < 0.2. \end{cases}$$

Both algorithms produce (in the long-run) the same probabilities, e.g.,  $\Pr(\text{whole life}) = 0.1$ , and so forth. So, neither is incorrect. You should be aware that there is more than one way to accomplish a goal. Similarly, you could use an alternative algorithm for ordered outcomes (such as failure times 1, 2, 3, 4, or 5, above).

### 8.1.3 Ready-made Random Number Generators

Sections 8.1.1 and 8.1.2 showed how one can generate simulated values from the foundations. This approach is important so that analyst can appreciate why the simulation works so well. However, because simulation is used so widely, it is not surprising that packages have been developed as time-saving devices.

For example, we have already seen in Example 8.1.3 that one can generate exponentially distributed random variates through the `rexp` function. This function means that analyst need not generate uniform random variates and then transform them using the inverse exponential distribution function. Instead, this is done in a single step using the `rexp` function.

**Table 8.2** summarizes a few of the standard random number generators in R; the `r` at the beginning of each function refers to a random number generator. Additional documentation for these functions are in Appendix Chapter ???. Note that the Pareto distribution requires the package `actuar`.

Table 8.2. Random Number Generators (RNGs)

Discrete Distributions		Continuous Distributions	
Distribution	RNG Function	Distribution	RNG Function
Binomial	<code>rbinom</code>	Exponential	<code>rexp</code>
Poisson	<code>rpoisson</code>	Gamma	<code>rgamma</code>
Negative Binomial	<code>rnbnom</code>	Pareto	<code>actuar::rpareto</code>
		Normal	<code>rnorm</code>
		Weibull	<code>rweibull</code>

### 8.1.4 Simulating from Complex Distributions

In statistical software programs such as R, analysts will find several ready-made random number generators. However, for many complex actuarial applications, it is likely that ready-made generators will not be available and so one must return to the foundations.

To illustrate, consider the aggregate claims distributions introduced in Chapter 7. There, in Section ??, we have already seen how to simulate aggregate loss distributions. As we saw in [Example 7.4.2], the process is to first simulate the number of losses and then simulate individual losses.

As another example of a complex distribution, consider the following example.

**Example 8.1.7. Generating Random Numbers from a Hybrid Distribution.** Consider a random variable that is 0 with probability 70% and is exponentially distributed with parameter  $\theta = 10,000$  with probability 30%. In an insurance application, this might correspond to a 70% chance of having no insurance claims and a 30% chance of a claim - if a claim occurs, then it is exponentially distributed. The distribution function, depicted in Figure 8.3, is given as

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - 0.3 \exp(-x/10000) & x \geq 0. \end{cases}$$

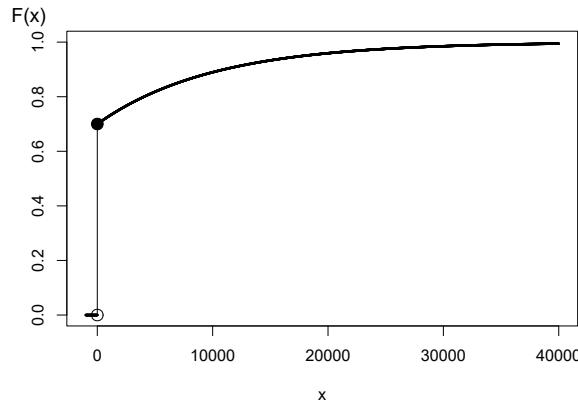


FIGURE 8.3: Distribution Function of a Hybrid Random Variable

From Figure 8.3, we can see that the inverse transform for generating random variables with this distribution function is

$$X = F^{-1}(U) = \begin{cases} 0 & 0 < U \leq 0.7 \\ -1000 \ln(\frac{1-U}{0.3}) & 0.7 < U < 1. \end{cases}$$

For discrete and hybrid random variables, the key is to draw a graph of the distribution function that allows you to visualize potential values of the inverse function.

You can think of this hybrid distribution as a special case of a mixture model that was introduced in Sections 3.6 and 4.3.2. Mixture models are straightforward to evaluate using simulation. In the first stage, one simulates a variable indicating the subpopulation. In the second stage, one simulates from that subpopulation. The resulting variate is a realization from the mixture model. To illustrate, let's revisit [Example 4.3.5](#).

**Example 4.3.5. Continued.** In this problem, we can label draws from the Type 1 subpopulation as  $X_1$  from an exponential distribution with mean 200, and those from the Type 2 subpopulation as  $X_2$  from a Pareto distribution with parameters  $\alpha = 3$  and  $\theta = 200$ . Here, 25% of policies are Type 1 and 75% of policies are Type 2.

We can use simulation to find the probability that the annual loss will be less than 100, and find the average loss. The illustrative code uses the ready-made random number generator functions `rbinom`, `rexp`, and `actuar::rpareto`.

```
nsim <- 100000
Z <- rbinom(nsim, prob = 0.75, size = 1)
X1 <- rexp(nsim, rate = 1/200)
X2 <- actuar::rpareto(nsim, shape = 3, scale = 200)
X <- (1 - Z) * X1 + Z * X2
# sum(X<100)/nsim mean(X)
```

### 8.1.5 Importance Sampling

Another class of important problems utilize distributions that are from a limited region. For example, when a loss has a deductible, the resulting claim represents the payment by an insurer that is not observed for amounts less than the deductible. This type of problem was considered extensively in Chapter 5. As another example, for claims that are extremely large, one may wish to restrict an analysis to only extremely large outcomes - discussions of *tails* of distributions will be taken up in Section ???. To address both types of problems, we now suppose that we wish to draw according to  $X$ , conditional on  $X \in [a, b]$ .

To this end, one can use an accept-reject mechanism : draw  $x$  from distribution  $F$

- if  $x \in [a, b]$  : keep it (“accept”)

- if  $x \notin [a, b]$  : draw another one (“reject”)

Observe that from  $n$  values initially generated, we keep here only  $[F(b) - F(a)] \cdot n$  draws, on average.

**Example 8.1.8. Draws from a Normal Distribution.** Suppose that we draw from a normal distribution with mean 2.5 and variance 1,  $N(2.5, 1)$ , but are only interested in draws greater than  $a = 2$  and less than  $b = 4$ . That is, we can only use  $F(4) - F(2) = \Phi(4 - 2.5) - \Phi(2 - 2.5) = 0.9332 - 0.3085 = 0.6247$  proportion of the draws. Figure 8.4 demonstrates that some draws lie with the interval  $(2, 4)$  and some are outside.

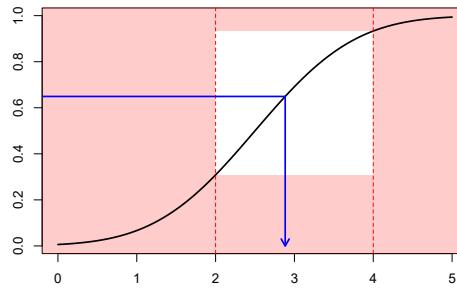


FIGURE 8.4: Demonstration of Draws In and Outside of  $(2,4)$

Instead, one can draw according to the conditional distribution  $F^*$  defined as

$$F^*(x) = \Pr(X \leq x | a < X \leq b) = \frac{F(x) - F(a)}{F(b) - F(a)}, \quad \text{for } a < x \leq b.$$

Using the inverse transform method in Section 8.1.2, we have that the draw

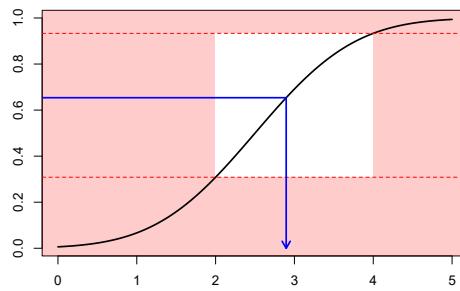
$$X^* = F^{*-1}(U) = F^{-1}(F(a) + U \cdot [F(b) - F(a)])$$

has distribution  $F^*$ . Expressed another way, define

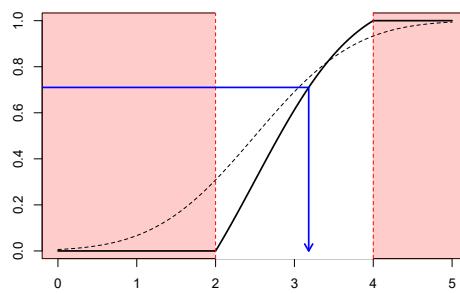
$$\tilde{U} = (1 - U) \cdot F(a) + U \cdot F(b)$$

and then use  $F^{-1}(\tilde{U})$ . With this approach, each draw counts.

This can be related to the importance sampling mechanism : we draw more frequently in regions where we expect to have quantities that have some interest. This transform can be considered as a “a change of measure.”



In Example 8.1.8, the inverse of the normal distribution is readily available (in R, the function is `qnorm`). However, for other applications, this is not always the case. Then, one simply uses numerical methods to determine  $X^*$  as the solution of the equation  $F(X^*) = \tilde{U}$  where  $\tilde{U} = (1 - U) \cdot F(a) + U \cdot F(b)$ .



## 8.2 Computing Distribution Parameters

In this section, you learn how to:

- Calculate quantities of interest and determine the precision of the calculated quantities
- Determine the appropriate number of replications for a simulation study
- Calculate complex distributions needed for hypothesis testing.

### 8.2.1 Simulating Parameters

One use of the term **parameter** is as a quantity that serves as an index for a known parametric family. For example, one usually thinks of a mean  $\mu$  and standard deviation  $\sigma$  as parameters of a normal distribution. Statisticians also use the term *parameter* to mean any quantity that summarizes a distribution. In this sense, a parameter can be written as  $\theta(F)$ , that is, if one knows the distribution function  $F(\cdot)$ , then one can compute the summary measure  $\theta$ .

In the previous subsection, we learned how to generate independent simulated realizations from a distribution of interest. With these realizations, we can construct an empirical distribution and approximate the underlying distribution as precisely as needed. As we introduce more actuarial applications in this book, you will see that simulation can be applied in a wide variety of contexts.

Many of these applications can be reduced to the problem of approximating a parameter  $E[h(X)]$ , where  $h(\cdot)$  is some known function. Based on  $R$  simulations (replications), we get  $X_1, \dots, X_R$ . From this simulated sample, we calculate an average

$$\bar{h}_R = \frac{1}{R} \sum_{i=1}^R h(X_i)$$

that we use as our simulated approximate (estimate) of  $E[h(X)]$ . To estimate the precision of this approximation, we use the simulation variance

$$s_{h,R}^2 = \frac{1}{R-1} \sum_{i=1}^R (h(X_i) - \bar{h}_R)^2.$$

From the independence, the standard error of the estimate is  $s_{h,R}/\sqrt{R}$ . This can be made as small as we like by increasing the number of replications  $R$ .

**Example 8.2.1. Portfolio Management.** In Section ??, we learned how to calculate the expected value of policies with deductibles. For an example of something that cannot be done with closed form expressions, we now consider two risks. This is a variation of a more complex example that will be covered as [Example 13.4.6](#).

We consider two property risks of a telecommunications firm:

- $X_1$  - buildings, modeled using a gamma distribution with mean 200 and scale parameter 100.
- $X_2$  - motor vehicles, modeled using a gamma distribution with mean 400 and scale parameter 200.

Denote the total risk as  $X = X_1 + X_2$ . For simplicity, you assume that these risks are independent.

To manage the risk, you seek some insurance protection. You are willing to retain internally small building and motor vehicles amounts, up to  $M$ , say. Random amounts in excess of  $M$  will have an unpredictable affect on your budget and so for these amounts you seek insurance protection. Stated mathematically, your retained risk is  $Y_{retained} = \min(X_1 + X_2, M)$  and the insurer's portion is  $Y_{insurer} = X - Y_{retained}$ .

To be specific, we use  $M = 400$  as well as  $R = 1000000$  simulations.

**a.** With these settings, we wish to determine the expected claim amount and the associated standard deviation of (i) that retained by you, (ii) that accepted by the insurer, and (iii) the total overall amount. **b.** For insured claims, the standard error of the simulation approximation is  $s_{h,R}/\sqrt{1000000} = 280.86/\sqrt{1000000} = 0.281$ . For this example, simulation is quick and so a large value such as 1000000 is an easy choice. However, for more complex problems, the simulation size may be an issue.

**Example Solution.** For part (a), the results of these calculations are:

	Retained	Insurer	Total
Mean	365.17	235.01	600.18
Standard Deviation	69.51	280.86	316.36

For part (b), Figure 8.5 allows us to visualize the development of the approximation as the number of simulations increases.

You can learn more about the R code for this example at the online version of this book, [Actuarial Community \(2025\)](#).

### 8.2.2 Determining the Number of Simulations

How many simulated values are recommended? 100? 1,000,000? We can use the central limit theorem to respond to this question.

As one criterion for your confidence in the result, suppose that you wish to be within 1% of the mean with 95% certainty. That is, you want  $\Pr(|\bar{h}_R - E[h(X)]| \leq 0.01E[h(X)]) \geq 0.95$ . According to the central limit theorem, your estimate should be approximately normally distributed and so we want to have  $R$  large enough to satisfy  $0.01E[h(X)]/\sqrt{\text{Var}[h(X)]/R} \geq 1.96$ . (Recall that 1.96 is the 97.5th percentile from the standard normal distribution.) Replacing  $E[h(X)]$  and  $\text{Var}[h(X)]$  with estimates, you continue

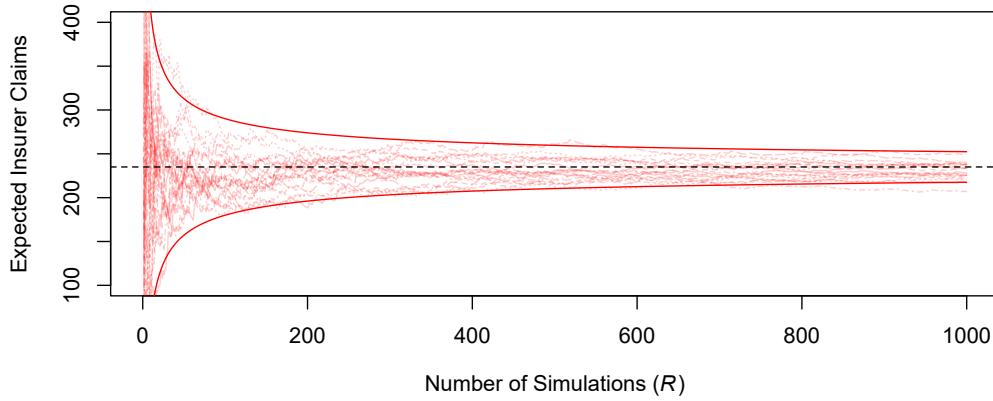


FIGURE 8.5: Estimated Expected Insurer Claims versus Number of Simulations

your simulation until

$$\frac{0.01 \bar{h}_R}{s_{h,R}/\sqrt{R}} \geq 1.96$$

or equivalently

$$R \geq 38,416 \frac{s_{h,R}^2}{\bar{h}_R^2}. \quad (8.1)$$

This criterion is a direct application of the approximate normality. Note that  $\bar{h}_R$  and  $s_{h,R}$  are not known in advance, so you will have to come up with estimates, either by doing a small pilot study in advance or by interrupting your procedure intermittently to see if the criterion is satisfied.

**Example 8.2.1. Portfolio Management - continued.** For this example, the average insurance claim is 235.011 and the corresponding standard deviation is 280.862. Using equation (8.1), to be within 1% of the mean, we would only require at least 54.87 thousand simulations. In addition, to be within 0.1% we would want at least 5.49 million simulations.

**Example 8.2.2. Approximation Choices.** An important application of simulation is the approximation of  $E [h(X)]$ . In this example, we show that the choice of the  $h(\cdot)$  function and the distribution of  $X$  can play a role.

Consider the following question : what is  $\Pr[X > 2]$  when  $X$  has a Cauchy distribution, with density  $f(x) = [\pi(1 + x^2)]^{-1}$ , on the real line? The true

value is

$$\Pr [X > 2] = \int_2^\infty \frac{dx}{\pi(1+x^2)}.$$

One can use an R numerical integration function (which usually works well on improper integrals)

which is equal to 0.14758.

**Approximation 1.** Alternatively, one can use simulation techniques to approximate that quantity. From calculus, you can check that the quantile function of the Cauchy distribution is  $F^{-1}(y) = \tan [\pi(y - 0.5)]$ . Then, with simulated uniform (0,1) variates,  $U_1, \dots, U_R$ , we can construct the estimator

$$p_1 = \frac{1}{R} \sum_{i=1}^R I(F^{-1}(U_i) > 2) = \frac{1}{R} \sum_{i=1}^R I(\tan [\pi(U_i - 0.5)] > 2).$$

With one million simulations, we obtain an estimate of 0.14744 with standard error 0.355 (divided by 1000). The estimated variance of  $p_1$  can be written as  $0.127/R$ .

**Approximation 2.** With other choices of  $h(\cdot)$  and  $F(\cdot)$  it is possible to reduce uncertainty even using the same number of simulations  $R$ . To begin, one can use the symmetry of the Cauchy distribution to write  $\Pr[X > 2] = 0.5 \cdot \Pr[|X| > 2]$ . With this, can construct a new estimator,

$$p_2 = \frac{1}{2R} \sum_{i=1}^R I(|F^{-1}(U_i)| > 2).$$

With one million simulations, we obtain an estimate of 0.14748 with standard error 0.228 (divided by 1000). The estimated variance of  $p_2$  can be written as  $0.052/R$ .

**Approximation 3.** But one can go one step further. The improper integral can be written as a proper one by a simple symmetry property (since the function is symmetric and the integral on the real line is equal to 1)

$$\int_2^\infty \frac{dx}{\pi(1+x^2)} = \frac{1}{2} - \int_0^2 \frac{dx}{\pi(1+x^2)}.$$

From this expression, a natural approximation would be

$$p_3 = \frac{1}{2} - \frac{1}{R} \sum_{i=1}^R h_3(2U_i), \quad \text{where } h_3(x) = \frac{2}{\pi(1+x^2)}.$$

With one million simulations, we obtain an estimate of 0.14756 with standard error 0.169 (divided by 1000). The estimated variance of  $p_3$  can be written as  $0.0285/R$ .

**Approximation 4.** Finally, one can also consider some change of variable in the integral

$$\int_2^\infty \frac{dx}{\pi(1+x^2)} = \int_0^{1/2} \frac{y^{-2}dy}{\pi(1-y^{-2})}.$$

From this expression, a natural approximation would be

$$p_4 = \frac{1}{R} \sum_{i=1}^R h_4(U_i/2), \quad \text{where } h_4(x) = \frac{1}{2\pi(1+x^2)}.$$

The expression seems rather similar to the previous one.

With one million simulations, we obtain an estimate of 0.14759 with standard error 0.01 (divided by 1000). The estimated variance of  $p_4$  can be written as  $0.00009/R$ , which is much smaller than what we had so far!

**Table 8.1** summarizes the four choices of  $h(\cdot)$  and  $F(\cdot)$  to approximate  $\Pr[X > 2] = 0.14758$ . The standard error varies dramatically. Thus, if we have a desired degree of accuracy, then the *number of simulations* depends strongly on how we write the integrals we try to approximate.

Table 8.1. Summary of Four Choices to Approximate  $\Pr[X > 2]$

Estimator	Definition	Support Function	Estimate	Standard Error
$p_1$	$\frac{1}{R} \sum_{i=1}^R \mathbf{I}(F^{-1}(U_i) > 2)$	$F^{-1}(u) = \tan(\pi(u - 0.5))$	0.147439	0.000355
$p_2$	$\frac{1}{2R} \sum_{i=1}^R \mathbf{I}( F^{-1}(U_i)  > 2)$	$F^{-1}(u) = \tan(\pi(u - 0.5))$	0.147477	0.000228
$p_3$	$\frac{1}{2} - \frac{1}{R} \sum_{i=1}^R h_3(2U_i)$	$h_3(x) = \frac{2}{\pi(1+x^2)}$	0.147558	0.000169
$p_4$	$\frac{1}{R} \sum_{i=1}^R h_4(U_i/2)$	$h_4(x) = \frac{1}{2\pi(1+x^2)}$	0.147587	0.000010

### 8.2.3 Simulation and Statistical Inference

Simulations not only help us approximate expected values but are also useful in calculating other aspects of distribution functions. As described in Section 8.2.1, the logic is that one wishes to calculate a parameter  $\theta(F)$ , use the same rule for calculating the parameter but replace the distribution function  $F(\cdot)$  with an empirical one from a simulated sample. For example, in addition to expected values, analysts can use simulation to compute quantiles from complex distributions.

In addition, simulation is very useful when distributions of test statistics are

too complicated to derive; in this case, one can use simulations to approximate the reference distribution. We now illustrate this with the Kolmogorov-Smirnov test which we learned about in Section ??.

**Example 8.2.3. Kolmogorov-Smirnov Test of Distribution.** Suppose that we have available  $n = 100$  observations  $\{x_1, \dots, x_n\}$  that, unknown to the analyst, were generated from a gamma distribution with parameters  $\alpha = 6$  and  $\theta = 2$ . The analyst believes that the data come from a lognormal distribution with parameters 1 and 0.4 and would like to test this assumption.

The first step is to visualize the data.

With this set-up, Figure 8.6 provides a graph of a histogram and empirical distribution. For reference, superimposed are red dashed lines from the lognormal distribution.

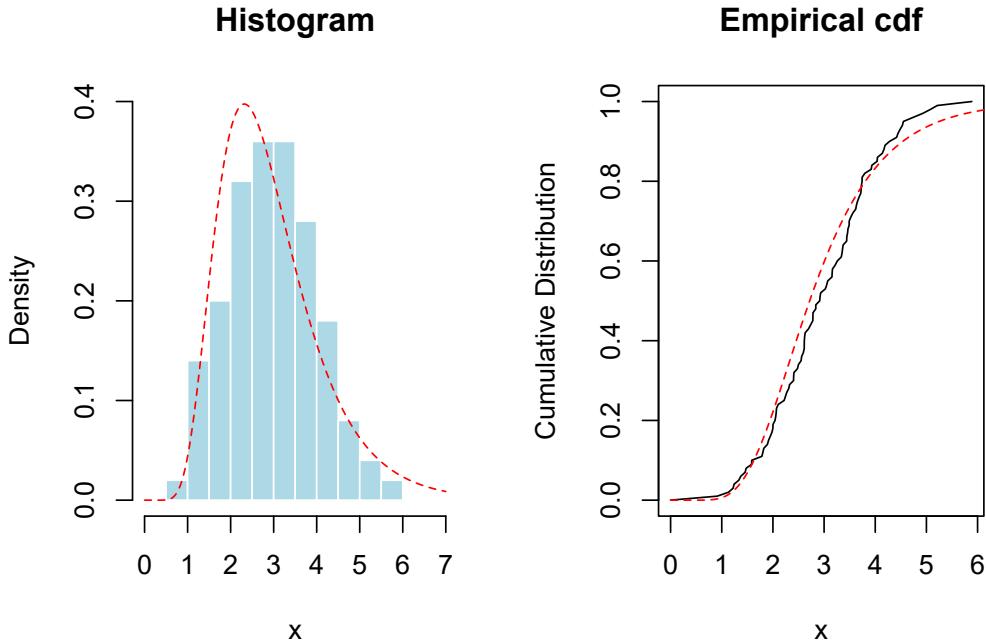


FIGURE 8.6: **Histogram and Empirical Distribution Function of Data used in Kolmogorov-Smirnov Test.** The red dashed lines are fits based on (incorrectly) hypothesized lognormal distribution.

Recall that the Kolmogorov-Smirnov statistic equals the largest discrepancy between the empirical and the hypothesized distribution. This is  $\max_x |F_n(x) - F_0(x)|$ , where  $F_0$  is the hypothesized lognormal distribution. We can calculate this directly.

Fortunately, for the lognormal distribution, R has built-in tests that allow us to determine this without complex programming:

```
ks.test(x, plnorm, mean = 1, sd = 0.4)
```

```
Asymptotic one-sample Kolmogorov-Smirnov test

data: x
D = 0.09703666, p-value = 0.303148
alternative hypothesis: two-sided
```

However, for many distributions of actuarial interest, pre-built programs are not available. We can use simulation to test the relevance of the test statistic. Specifically, to compute the  $p$ -value, let us generate thousands of random samples from a  $LN(1, 0.4)$  distribution (with the same size), and compute empirically the distribution of the statistic,

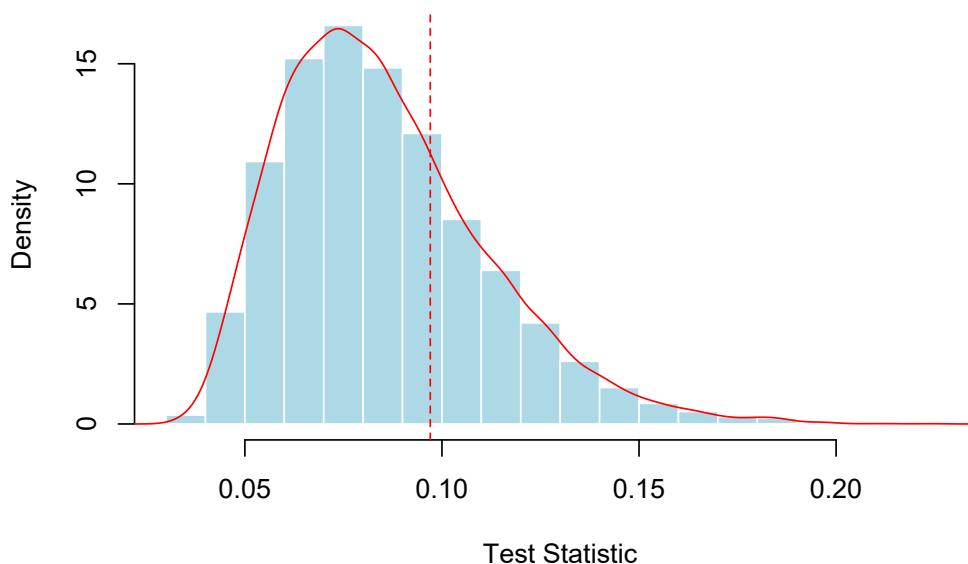
```
ns <- 10000
d_KS <- rep(NA, ns)
# compute the test statistics for a large (ns) number of simulated samples
for (s in 1:ns) d_KS[s] <- D(rlnorm(n, 1, 0.4), function(x) plnorm(x, 1, 0.4))
mean(d_KS > D(x, function(x) plnorm(x, 1, 0.4)))
```

```
[1] 0.2843
```

The simulated distribution based on 10,000 random samples is summarized in Figure 8.7. Here, the statistic exceeded the empirical value (0.09704) in 28.43% of the scenarios, while the *theoretical*  $p$ -value is 0.3031. For both the simulation and the theoretical  $p$ -values, the conclusions are the same; the data do not provide sufficient evidence to reject the hypothesis of a lognormal distribution.

Although only an approximation, the simulation approach works in a variety of distributions and test statistics without needing to develop the nuances of the underpinning theory for each situation. We summarize the procedure for developing simulated distributions and  $p$ -values as follows:

1. Draw a sample of size  $n$ , say,  $X_1, \dots, X_n$ , from a known distribution function  $F$ . Compute a statistic of interest, denoted as  $\hat{\theta}(X_1, \dots, X_n)$ . Call this  $\hat{\theta}^r$  for the  $r$ th replication.
2. Repeat this  $r = 1, \dots, R$  times to get a sample of statistics,  $\hat{\theta}^1, \dots, \hat{\theta}^R$ .
3. From the sample of statistics in Step 2,  $\{\hat{\theta}^1, \dots, \hat{\theta}^R\}$ , compute a summary measure of interest, such as a  $p$ -value.



**FIGURE 8.7: Simulated Distribution of the Kolmogorov-Smirnov Test Statistic.** The vertical red dashed line marks the test statistic for the sample of 100.

---

## 8.3 Bootstrapping and Resampling

---

In this section, you learn how to:

- Generate a nonparametric bootstrap distribution for a statistic of interest
  - Use the bootstrap distribution to generate estimates of precision for the statistic of interest, including bias, standard deviations, and confidence intervals
  - Perform bootstrap analyses for parametric distributions
- 

### 8.3.1 Bootstrap Foundations

Simulation presented up to now is based on sampling from a **known** distribution. Section 8.1 showed how to use simulation techniques to sample and compute quantities from known distributions. However, statistical science is dedicated to providing inferences about distributions that are *unknown*. We gather summary statistics based on this unknown population distribution. But how do we sample from an unknown distribution?

Naturally, we cannot simulate draws from an unknown distribution but we can draw from a sample of observations. If the sample is a good representation from the population, then our simulated draws from the sample should well approximate the simulated draws from a population. The process of sampling from a sample is called *resampling* or *bootstrapping*. The term bootstrap comes from the phrase “pulling oneself up by one’s bootstraps” Efron (1979). With resampling, the original sample plays the role of the population and estimates from the sample play the role of true population parameters.

The resampling algorithm is the same as introduced in Section 8.2.3 except that now we use simulated draws from a sample. It is common to use  $\{X_1, \dots, X_n\}$  to denote the original sample and let  $\{X_1^*, \dots, X_n^*\}$  denote the simulated draws. We draw them with replacement so that the simulated draws will be independent from one another, the same assumption as with the original sample. For each sample, we also use  $n$  simulated draws, the same number as the original sample size. To distinguish this procedure from the simulation, it is common to use  $B$  (for bootstrap) to be the number of simulated samples. We could also write  $\{X_1^{(b)}, \dots, X_n^{(b)}\}$ ,  $b = 1, \dots, B$  to clarify this.

There are two basic resampling methods, *model-free* and *model-based*, which are, respectively, as *nonparametric* and *parametric*. In the nonparametric ap-

proach, no assumption is made about the distribution of the parent population. The simulated draws come from the empirical distribution function  $F_n(\cdot)$ , so each draw comes from  $\{X_1, \dots, X_n\}$  with probability  $1/n$ .

In contrast, for the parametric approach, we assume that we have knowledge of the distribution family  $F$ . The original sample  $X_1, \dots, X_n$  is used to estimate parameters of that family, say,  $\hat{\theta}$ . Then, simulated draws are taken from the  $F(\hat{\theta})$ . Section 8.3.4 discusses this approach in further detail.

### Nonparametric Bootstrap

The idea of the nonparametric bootstrap is to use the inverse transform method on  $F_n$ , the empirical cumulative distribution function, depicted in Figure 8.8.

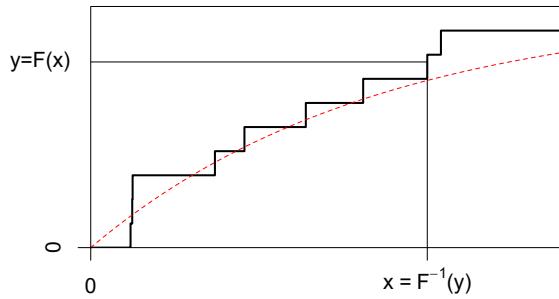


FIGURE 8.8: Inverse of an Empirical Distribution Function

Because  $F_n$  is a step-function,  $F_n^{-1}$  takes values in  $\{x_1, \dots, x_n\}$ . More precisely, as illustrated in Figure 8.9.

- if  $y \in (0, 1/n)$  (with probability  $1/n$ ) we draw the smallest value ( $\min\{x_i\}$ )
- if  $y \in (1/n, 2/n)$  (with probability  $1/n$ ) we draw the second smallest value,
- $\vdots \vdots \vdots$
- if  $y \in ((n-1)/n, 1)$  (with probability  $1/n$ ) we draw the largest value ( $\max\{x_i\}$ ).

Using the inverse transform method with  $F_n$  means sampling from  $\{x_1, \dots, x_n\}$ , with probability  $1/n$ . Generating a bootstrap sample of size  $B$  means sampling from  $\{x_1, \dots, x_n\}$ , with probability  $1/n$ , with replacement. See the following illustrative R code.

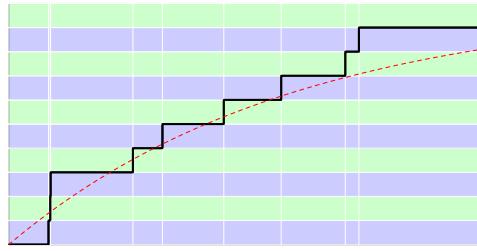


FIGURE 8.9: Inverse of an Empirical Distribution Function

```
set.seed(1)
n <- 10
x <- rexp(n, 1/6)
m <- 10
bootvalues <- sample(x, size = m, replace = TRUE)
```

```
[1] 2.6164 5.7394 5.7394 2.6164 2.6164 7.0899 0.8823 5.7394 4.5311 0.8388
```

Observe that value 5.7394 was obtained three times.

### 8.3.2 Bootstrap Precision: Bias, Standard Deviation, and Mean Square Error

We summarize the nonparametric bootstrap procedure as follows:

1. From the sample  $\{X_1, \dots, X_n\}$ , draw a sample of size  $n$  (with replacement), say,  $X_1^*, \dots, X_n^*$ . From the simulated draws compute a statistic of interest, denoted as  $\hat{\theta}(X_1^*, \dots, X_n^*)$ . Call this  $\hat{\theta}_b^*$  for the  $b$ th replicate.
2. Repeat this  $b = 1, \dots, B$  times to get a sample of statistics,  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ .
3. From the sample of statistics in Step 2,  $\{\hat{\theta}_1^*, \dots, \hat{\theta}_B^*\}$ , compute a summary measure of interest.

In this section, we focus on three summary measures, the bias, the standard deviation, and the mean square error (*MSE*). **Table 8.3** summarizes these three measures. Here,  $\bar{\hat{\theta}}^*$  is the average of  $\{\hat{\theta}_1^*, \dots, \hat{\theta}_B^*\}$ .

Table 8.3. Bootstrap Summary Measures

<i>Population Measure</i>	<i>Population Definition</i>	<i>Bootstrap Approximation</i>	<i>Bootstrap Symbol</i>
Bias	$E(\hat{\theta}) - \theta$	$\bar{\hat{\theta}}^* - \hat{\theta}$	$Bias_{boot}(\hat{\theta})$
Standard Deviation	$\sqrt{Var(\hat{\theta})}$	$\sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b^* - \bar{\hat{\theta}}^*)^2}$	$s_{boot}(\hat{\theta})$
Mean Square Error	$E(\hat{\theta} - \theta)^2$	$\frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta})^2$	$MSE_{boot}(\hat{\theta})$

**Example 8.3.1. Bodily Injury Claims and Loss Elimination Ratios.**

To show how the bootstrap can be used to quantify the precision of estimators, we return to the [Example 5.3.2](#) bodily injury claims data where we introduced a nonparametric estimator of the loss elimination ratio.

[Table 8.4](#) summarizes the results of the bootstrap estimation. For example, at  $d = 14000$ , the nonparametric estimate of *LER* is 0.97678. This has an estimated bias of 0.00016 with a standard deviation of 0.00687. For some applications, you may wish to apply the estimated bias to the original estimate to give a bias-corrected estimator. This is the focus of the next example. For this illustration, the bias is small and so such a correction is not relevant.

Table 8.4. **Bootstrap Estimates of LER at Selected Deductibles**

d	NP	Bootstrap		Lower Normal		Upper Normal	
		Estimate	Bias	SD	95% CI	95% CI	
4000	0.54113	0.00011	0.01237		0.51678	0.56527	
5000	0.64960	0.00027	0.01412		0.62166	0.67700	
10500	0.93563	0.00004	0.01017		0.91567	0.95553	
11500	0.95281	-0.00003	0.00941		0.93439	0.97128	
14000	0.97678	0.00016	0.00687		0.96316	0.99008	
18500	0.99382	0.00014	0.00331		0.98719	1.00017	

The bootstrap standard deviation gives a measure of precision. For one application of standard deviations, we can use the normal approximation to create a confidence interval. For example, the R function `boot.ci` produces the normal confidence intervals at 95%. These are produced by creating an interval of twice the length of 1.95994 bootstrap standard deviations, centered about the bias-corrected estimator (1.95994 is the 97.5th quantile of the standard normal distribution). For example, the lower normal 95% CI at  $d = 14000$  is  $(0.97678 - 0.00016) - 1.95994 \times 0.00687 = 0.96316$ . We further discuss bootstrap confidence intervals in the next section.

**Example 8.3.2. Estimating  $\log(\mu)$ .** The bootstrap can be used to quantify the bias of an estimator, for instance. Consider here a sample  $\mathbf{x} = \{x_1, \dots, x_n\}$  that is iid with mean  $\mu$ .

```
sample_x <- c(2.46, 2.8, 3.28, 3.86, 2.85, 3.67, 3.37, 3.4, 5.22, 2.55, 2.79, 4.5,
3.37, 2.88, 1.44, 2.56, 2, 2.07, 2.19, 1.77)
```

Suppose that the quantity of interest is  $\theta = \log(\mu)$ . A natural estimator would be  $\hat{\theta}_1 = \log(\bar{x})$ . This estimator is biased (due to the Jensen inequality) but is asymptotically unbiased. For our sample, the estimate is as follows.

```
(theta_1 <- log(mean(sample_x)))
```

```
[1] 1.08231352
```

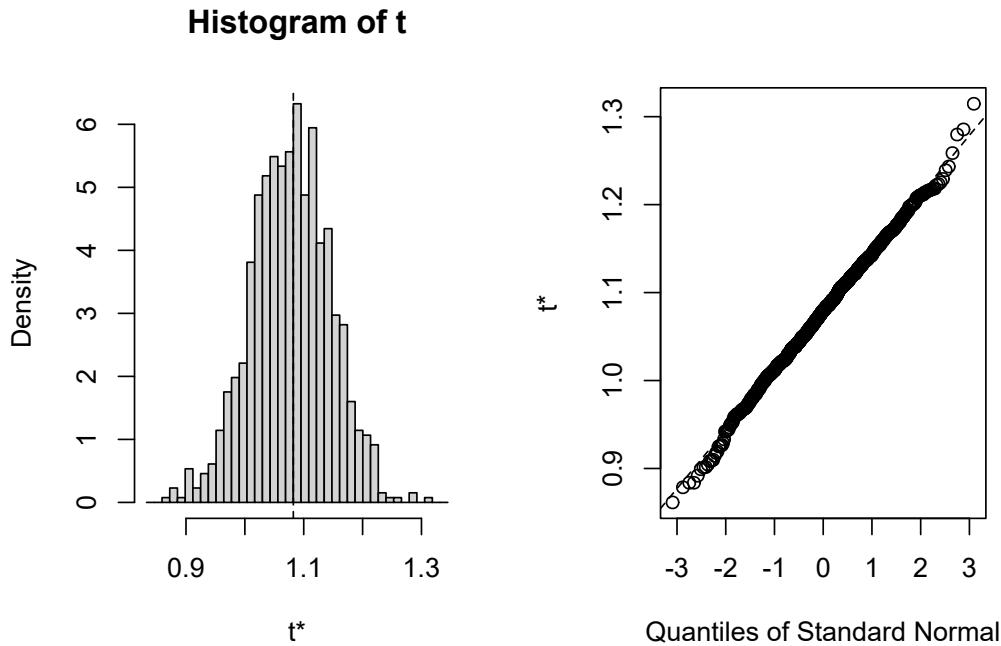
One can use a bootstrap strategy to get a correction: given a bootstrap sample,  $\mathbf{x}_b^*$ , let  $\bar{x}_b^*$  denote its mean, and set

$$\hat{\theta}_2 = \frac{1}{B} \sum_{b=1}^B \log(\bar{x}_b^*).$$

To implement this, we have the following code where we now use the function `boot` from the R package `boot`.

```
library(boot)
results <- boot(data = sample_x, statistic = function(y, indices) {
  log(mean(y[indices]))
}, R = 1000)
theta_2 <- 2 * theta_1 - mean(results$t)
```

Then, you can `plot(results)` and `print(results)` to see the following.



**FIGURE 8.10: Distribution of Bootstrap Replicates.** The left-hand panel is a histogram of replicates. The right-hand panel is a quantile-quantile plot, comparing the bootstrap distribution to the standard normal distribution.

#### ORDINARY NONPARAMETRIC BOOTSTRAP

```
Call:
boot(data = sample_x, statistic = function(y, indices) {
  log(mean(y[indices]))
}, R = 1000)
```

```
Bootstrap Statistics :
original      bias    std. error
t1* 1.08231352 -0.00438957075 0.0669312212
```

This results in two estimators, the raw estimator  $\hat{\theta}_1 = 1.082$  and the bootstrap estimator  $\hat{\theta}_2 = 1.087$ .

How does this work with differing sample sizes? We now suppose that the  $x_i$ 's are generated from a gamma distribution with shape parameter  $\alpha = 0.25$  and scale parameter  $\theta = 12$ . We use simulation to draw the sample sizes but then

act as if they were a realized set of observations. See the following illustrative code.

```

param <- function(x) {
  n <- length(x)
  theta_1 <- log(mean(x))
  results <- boot(data = x, statistic = function(y, indices) {
    log(mean(y[indices]))
  }, R = 999)
  theta_2 <- 2 * theta_1 - mean(results$t)
  return(c(theta_1, theta_2))
}
set.seed(2074)
ns <- 200
est <- function(n) {
  call_param <- function(i) {
    param(rgamma(n, shape = 0.25, scale = 12))
  }
  V <- Vectorize(call_param)(1:ns)
  apply(V, 1, median)
}
VN <- seq(15, 100, by = 5)
Est <- Vectorize(est)(VN)

save(VN, Est, file = "../IntermediateCalcs/SimulationChapter/Section832Bootstrap.Rdata")

```

The results of the comparison are summarized in Figure 8.11. This figure shows that the bootstrap estimator is closer to the true parameter value for many of the sample sizes. The bias of both estimators decreases as the sample size increases.

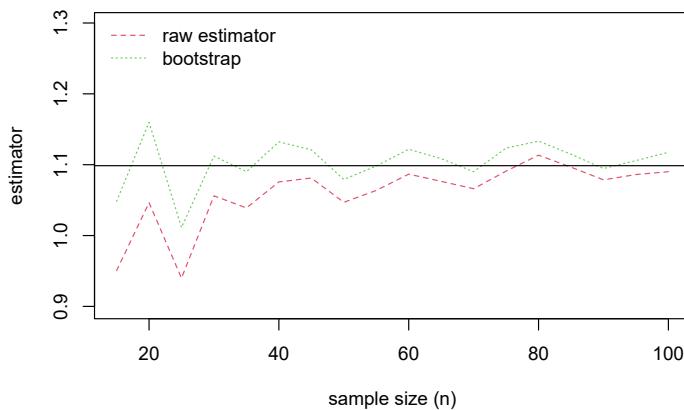


FIGURE 8.11: **Comparison of Estimates.** True value of the parameter is given by the solid horizontal line at  $\log(3) \approx 1.099$ .

Although successful in this example, we remark that the bootstrap bias adjusted estimator is generally not used in practice because the bias adjustment introduces extra variability into the estimator. Instead, the bias estimate provides information as to whether or not the estimate contains bias; this information gives additional information about the reliability of the estimate.

### 8.3.3 Confidence Intervals

The bootstrap procedure generates  $B$  replicates  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$  of the estimator  $\hat{\theta}$ . In [Example 8.3.1](#), we saw how to use standard normal approximations to create a confidence interval for parameters of interest. However, given that a major point is to use bootstrapping to avoid relying on assumptions of approximate normality, it is not surprising that there are alternative confidence intervals available.

For an estimator  $\hat{\theta}$ , the *basic* bootstrap confidence interval is

$$(2\hat{\theta} - q_U, 2\hat{\theta} - q_L), \quad (8.2)$$

where  $q_L$  and  $q_U$  are lower and upper 2.5% quantiles from the bootstrap sample  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ .

To see where this comes from, start with the idea that  $(q_L, q_U)$  provides a 95% interval for  $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ . So, for a random  $\hat{\theta}_b^*$ , there is a 95% chance that  $q_L \leq \hat{\theta}_b^* \leq q_U$ . Reversing the inequalities and adding  $\hat{\theta}$  to each side gives a 95% interval

$$\hat{\theta} - q_U \leq \hat{\theta} - \hat{\theta}_b^* \leq \hat{\theta} - q_L.$$

So,  $(\hat{\theta} - q_U, \hat{\theta} - q_L)$  is an 95% interval for  $\hat{\theta} - \hat{\theta}_b^*$ . The bootstrap approximation idea says that this is also a 95% interval for  $\theta - \hat{\theta}$ . Adding  $\hat{\theta}$  to each side gives the 95% interval in equation (8.2).

Many alternative bootstrap intervals are available. The easiest to explain is the percentile bootstrap interval which is defined as  $(q_L, q_U)$ . However, this has the drawback of potentially poor behavior in the tails which can be of concern in some actuarial problems of interest.

**Example 8.3.3. Bodily Injury Claims and Risk Measures.** To see how the bootstrap confidence intervals work, we return to the bodily injury auto claims considered in [Example 8.3.1](#). Instead of the loss elimination ratio, suppose we wish to estimate the 95th percentile  $F^{-1}(0.95)$  and a measure defined as

$$ES_{0.95}[X] = E[X|X > F^{-1}(0.95)].$$

This measure is called the expected shortfall. In this formulation, it is the expected value of  $X$  conditional on  $X$  exceeding the 95th percentile which is also sometimes known as the *conditional value at risk*. Section ?? explains how quantiles and the expected shortfall are the two most important examples of so-called *risk measures*. For now, we will simply think of these as measures that we wish to estimate. For the percentile, we use the nonparametric estimator  $F_n^{-1}(0.95)$  defined in Section 4.4.1. For the expected shortfall, we use the plug-in principle to define the nonparametric estimator

$$ES_{n,0.95}[X] = \frac{\sum_{i=1}^n X_i I[X_i > F_n^{-1}(0.95)]}{\sum_{i=1}^n I[X_i > F_n^{-1}(0.95)]}.$$

In this expression, the denominator counts the number of observations that exceed the 95th percentile  $F_n^{-1}(0.95)$ . The numerator adds up losses for those observations that exceed  $F_n^{-1}(0.95)$ . Table 8.5 summarizes the estimator for selected fractions.

**Table 8.5. Bootstrap Estimates of Quantiles at Selected Fractions**

Fraction	NP	Bootstrap	Bootstrap	Lower Normal	Upper Normal	Lower Basic	Upper Basic	Lower Percentile	Upper Percentile
	Estimate	Bias	SD	95% CI	95% CI	95% CI	95% CI	95% CI	95% CI
0.50	6500.00	-128.02	200.36	6235.32	7020.72	6300.00	7000.00	6000.00	6700.00
0.80	9078.40	89.51	200.27	8596.38	9381.41	8533.20	9230.40	8926.40	9623.60
0.90	11454.00	55.95	480.66	10455.96	12340.13	10530.49	12415.00	10493.00	12377.51
0.95	13313.40	13.59	667.74	11991.07	14608.55	11509.70	14321.00	12305.80	15117.10
0.98	16758.72	101.46	1273.45	14161.34	19153.19	14517.44	19326.95	14190.49	19000.00

For example, when the fraction is 0.50, we see that lower and upper 2.5th quantiles of the bootstrap simulations are  $q_L = 6000$  and  $q_u = 6700$ , respectively. These form the percentile bootstrap confidence interval. With the nonparametric estimator 6500, these yield the lower and upper bounds of the basic confidence interval 6300 and 7000, respectively. Table 8.5 also shows bootstrap estimates of the bias, standard deviation, and a normal confidence interval, concepts introduced in Section 8.3.2.

Table 8.6 shows similar calculations for the expected shortfall. In each case, we see that the bootstrap standard deviation increases as the fraction increases. This is because there are fewer observations to estimate quantiles as the fraction increases, leading to greater imprecision. Confidence intervals also become wider. Interestingly, there does not seem to be the same pattern in the estimates of the bias.

Table 8.6. Bootstrap Estimates of ES at Selected Risk Levels

Fraction	NP	Bootstrap	Bootstrap	Lower Normal	Upper Normal	Lower Basic	Upper Basic	Lower Percentile	Upper Percentile
	Estimate	Bias	SD	95% CI	95% CI	95% CI	95% CI	95% CI	95% CI
0.50	9794.69	-120.82	273.35	9379.74	10451.27	9355.14	10448.87	9140.51	10234.24
0.80	12454.18	30.68	481.88	11479.03	13367.96	11490.62	13378.52	11529.84	13417.74
0.90	14720.05	17.51	718.23	13294.82	16110.25	13255.45	16040.72	13399.38	16184.65
0.95	17072.43	5.99	1103.14	14904.31	19228.56	14924.50	19100.88	15043.97	19220.36
0.98	20140.56	73.43	1587.64	16955.40	23178.85	16942.36	22984.40	17296.71	23338.75

### 8.3.4 Parametric Bootstrap

The idea of the nonparametric bootstrap is to resample by drawing independent variables from the empirical cumulative distribution function  $F_n$ . In contrast, with parametric bootstrap, we draw independent variables from  $F_{\hat{\theta}}$  where the underlying distribution is assumed to be in a parametric family such as a gamma or lognormal distribution. Typically, parameters from this distribution are estimated based on a sample and denoted as  $\hat{\theta}$ .

**Example 8.3.4. Lognormal distribution.** Consider again the dataset

```
sample_x <- c(2.46, 2.8, 3.28, 3.86, 2.85, 3.67, 3.37, 3.4, 5.22, 2.55, 2.79, 4.5,
            3.37, 2.88, 1.44, 2.56, 2, 2.07, 2.19, 1.77)
```

The classical (nonparametric) bootstrap was based on the following samples.

```
x <- sample(sample_x, replace = TRUE)
```

Instead, for the parametric bootstrap, we have to assume that the distribution of  $x_i$ 's is from a specific family. As an example, the following code utilizes a lognormal distribution.

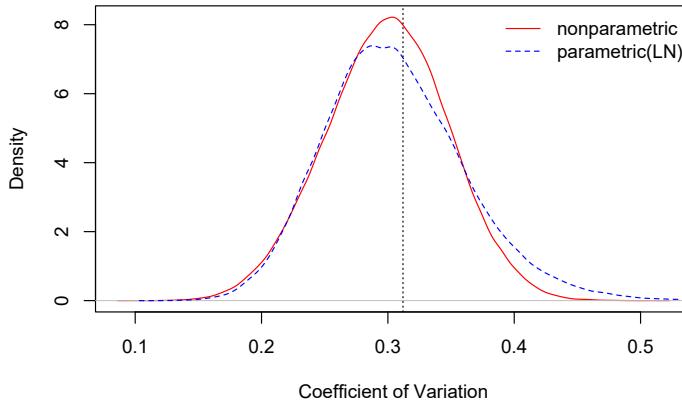
```
library(MASS)
fit <- fitdistr(sample_x, dlnorm, list(meanlog = 1, sdlog = 1))
fit
```

```
meanlog      sdlog
1.0363069735  0.3059343996
(0.0684090114) (0.0483702729)
```

Then we draw from that distribution.

```
x <- rlnorm(length(sample_x), meanlog = fit$estimate[1], sdlog = fit$estimate[2])
```

Figure 8.12 compares the bootstrap distributions for the coefficient of variation, one based on the nonparametric approach and the other based on a parametric approach, assuming a lognormal distribution.



**FIGURE 8.12: Comparison of Nonparametric and Parametric Bootstrap Distributions for the Coefficient of Variation**

**Example 8.3.5. Bootstrapping Censored Observations.** The parametric bootstrap draws simulated realizations from a parametric estimate of the distribution function. In the same way, we can draw simulated realizations from estimates of a distribution function. As one example, we might draw from smoothed estimates of a distribution function introduced in Section 4.4.1. Another special case, considered here, is to draw an estimate from the Kaplan-Meier estimator introduced in Section 5.3.3. In this way, we can handle observations that are censored.

Specifically, return to the bodily injury data in Examples 8.2.1 and 8.2.3 but now we include the 17 claims that were censored by policy limits. In Example 4.3.6, we used this full dataset to estimate the Kaplan-Meier estimator of the survival function introduced in Section 5.3.3. Table 8.7 presents bootstrap estimates of the quantiles from the Kaplan-Meier survival function estimator. These include the bootstrap precision estimates, bias and standard deviation, as well as the basic 95% confidence interval.

**Table 8.7. Bootstrap Kaplan-Meier Estimates of Quantiles at Selected Fractions**

Fraction	KM NP	Bootstrap	Bootstrap	Lower Basic	Upper Basic
	Estimate	Bias	SD	95% CI	95% CI
0.50	6500	18.77	177.38	6067	6869
0.80	9500	167.08	429.59	8355	9949
0.90	12756	37.73	675.21	10812	13677
0.95	18500	Inf	NaN	12500	22300
0.98	25000	Inf	NaN	-Inf	27308

Results in [Table 8.7](#) are consistent with the results for the uncensored subsample in [Table 8.5](#). In [Table 8.7](#), we note the difficulty in estimating quantiles at large fractions due to the censoring. However, for moderate size fractions (0.50, 0.80, and 0.90), the Kaplan-Meier nonparametric (KM NP) estimates of the quantile are consistent with those [Table 8.5](#). The bootstrap standard deviation is smaller at the 0.50 (corresponding to the median) but larger at the 0.80 and 0.90 levels. The censored data analysis summarized in [Table 8.7](#) uses more data than the uncensored subsample analysis in [Table 8.5](#) but also has difficulty extracting information for large quantiles.

## 8.4 Model Selection and Cross-Validation

In this section, you learn how to:

- Compare and contrast cross-validation to simulation techniques and bootstrap methods.
- Use cross-validation techniques for model selection
- Explain the jackknife method as a special case of cross-validation and calculate jackknife estimates of bias and standard errors

Cross-validation, briefly introduced in Chapter 2 and Section ??, is a technique based on simulated outcomes that is especially useful for selecting an appropriate model. We now compare and contrast cross-validation to other simulation techniques already introduced in this chapter.

- Simulation, or Monte-Carlo, introduced in Section 8.1, allows us to compute expected values and other summaries of statistical distributions, such as  $p$ -values, readily.

- Bootstrap, and other resampling methods introduced in Section 8.3, provides estimators of the precision, or variability, of statistics.
- Cross-validation is important when assessing how accurately a predictive model will perform in practice.

Overlap exists but nonetheless it is helpful to think about the broad goals associated with each statistical method.

To discuss cross-validation, let us recall from Chapter 2 some of the key ideas of model validation. When assessing, or validating, a model, we look to performance measured on *new* data, or at least not those that were used to fit the model. A classical approach is to split the sample in two: a subpart (the *training* dataset) is used to fit the model and the other one (the *testing* dataset) is used to validate. However, a limitation of this approach is that results depend on the split; even though the overall sample is fixed, the split between training and test subsamples varies randomly. A different training sample means that model estimated parameters will differ. Different model parameters and a different test sample means that validation statistics will differ. Two analysts may use the same data and same models yet reach different conclusions about the viability of a model (based on different random splits), a frustrating situation.

#### 8.4.1 k-Fold Cross-Validation

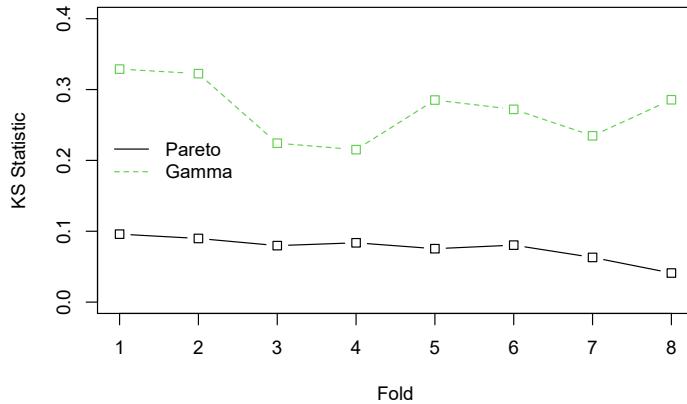
To mitigate this difficulty, it is common to use a cross-validation approach as introduced in Section 4.2.4. The key idea is to emulate the basic test/training approach to model validation by repeating it many times through averaging over different splits of the data. A key advantage is that the validation statistic is not tied to a specific parametric (or nonparametric) model - one can use a nonparametric statistic or a statistic that has economic interpretations - and so this can be used to compare models that are not nested (unlike likelihood ratio procedures).

**Example 8.4.1. Wisconsin Property Fund.** For the 2010 property fund data introduced in Section 1.3, we fit gamma and Pareto distributions to the 1,377 claims data. For details of the related goodness of fit, see Appendix Section 15.4.4. We now consider the Kolmogorov-Smirnov statistic introduced in Section ???. When the entire dataset was fit, the Kolmogorov-Smirnov goodness of fit statistic for the gamma distribution turns out to be 0.2639 and for the Pareto distribution is 0.0478. The lower value for the Pareto distribution indicates that this distribution is a better fit than the gamma.

To see how k-fold cross-validation works, we randomly split the data into  $k = 8$  groups, or folds, each having about  $1377/8 \approx 172$  observations. Then, we fit

gamma and Pareto models to a data set with the first seven folds (about  $172 \times 7 = 1,204$  observations), determine estimated parameters, and then used these fitted models with the held-out data to determine the Kolmogorov-Smirnov statistic.

The results appear in Figure 8.13 where horizontal axis is Fold=1. This process was repeated for the other seven folds. The results summarized in Figure 8.13 show that the Pareto consistently provides a more reliable predictive distribution than the gamma.



**FIGURE 8.13: Cross Validated Kolmogorov-Smirnov (KS) Statistics for the Property Fund Claims Data.** The solid black line is for the Pareto distribution, the green dashed line is for the gamma distribution. The KS statistic measures the largest deviation between the fitted distribution and the empirical distribution for each of 8 groups, or folds, of randomly selected data.

#### 8.4.2 Leave-One-Out Cross-Validation

A special case where  $k = n$  is known as leave-one-out cross validation. This case is historically prominent and is closely related to jackknife statistics, a precursor of the bootstrap technique.

Even though we present it as a special case of cross-validation, it is helpful to give an explicit definition. Consider a generic statistic  $\hat{\theta} = t(x)$  that is an estimator for a parameter of interest  $\theta$ . The idea of the jackknife is to compute  $n$  values  $\hat{\theta}_{-i} = t(x_{-i})$ , where  $x_{-i}$  is the subsample of  $x$  with the  $i$ -th value removed. The average of these values is denoted as

$$\bar{\hat{\theta}}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{-i}$$

These values can be used to create estimates of the bias of the statistic  $\hat{\theta}$

$$Bias_{jack} = (n - 1) \left( \bar{\hat{\theta}}_{(\cdot)} - \hat{\theta} \right) \quad (8.3)$$

as well as a standard deviation estimate

$$s_{jack} = \sqrt{\frac{n-1}{n} \sum_{i=1}^n \left( \hat{\theta}_{-i} - \bar{\hat{\theta}}_{(\cdot)} \right)^2}. \quad (8.4)$$

**Example 8.4.2. Coefficient of Variation.** To illustrate, consider a small fictitious sample  $x = \{x_1, \dots, x_n\}$  with realizations

```
sample_x <- c(2.46, 2.8, 3.28, 3.86, 2.85, 3.67, 3.37, 3.4, 5.22, 2.55, 2.79, 4.5,
3.37, 2.88, 1.44, 2.56, 2, 2.07, 2.19, 1.77)
```

Suppose that we are interested in the coefficient of variation  $\theta = CV = \sqrt{\text{Var}[X]/E[X]}$ .

With this dataset, the estimator of the coefficient of variation turns out to be 0.31196. But how reliable is it? To answer this question, we can compute the jackknife estimates of bias and its standard deviation. The following code shows that the jackknife estimator of the bias is  $Bias_{jack} = -0.00627$  and the jackknife standard deviation is  $s_{jack} = 0.01293$ .

```
# Sample Code for Example 8.4.2
CVar <- function(x) sqrt(var(x))/mean(x)
JackCVar <- function(i) sqrt(var(sample_x[-i]))/mean(sample_x[-i])
JackTheta <- Vectorize(JackCVar)(1:length(sample_x))
BiasJack <- (length(sample_x) - 1) * (mean(JackTheta) - CVar(sample_x))
sdJack <- sd(JackTheta)
```

**Example 8.4.3. Bodily Injury Claims and Loss Elimination Ratios.** In [Example 8.3.1](#), we showed how to compute bootstrap estimates of the bias and standard deviation for the loss elimination ratio using the [Example 4.1.11](#) bodily injury claims data. We follow up now by providing comparable quantities using jackknife statistics.

[Table 8.8](#) summarizes the results of the jackknife estimation. It shows that jackknife estimates of the bias and standard deviation of the loss elimination

ratio  $E[\min(X, d)]/E[X]$  are largely consistent with the bootstrap methodology. Moreover, one can use the standard deviations to construct normal based confidence intervals, centered around a bias-corrected estimator. For example, at  $d = 14000$ , we saw in Example 4.1.11 that the nonparametric estimate of  $LER$  is 0.97678. This has an estimated bias of 0.00010, resulting in the (jackknife) *bias-corrected* estimator 0.97688. The 95% confidence intervals are produced by creating an interval of twice the length of 1.96 jackknife standard deviations, centered about the bias-corrected estimator (1.96 is the approximate 97.5th quantile of the standard normal distribution).

Table 8.8. Jackknife Estimates of LER at Selected Deductibles

d	NP	Bootstrap	Bootstrap	Jackknife	Jackknife	Lower Jackknife	Upper Jackknife
	Estimate	Bias	SD	Bias	SD	95% CI	95% CI
4000	0.54113	0.00011	0.01237	0.00031	0.00061	0.53993	0.54233
5000	0.64960	0.00027	0.01412	0.00033	0.00068	0.64825	0.65094
10500	0.93563	0.00004	0.01017	0.00019	0.00053	0.93460	0.93667
11500	0.95281	-0.00003	0.00941	0.00016	0.00047	0.95189	0.95373
14000	0.97678	0.00016	0.00687	0.00010	0.00034	0.97612	0.97745
18500	0.99382	0.00014	0.00331	0.00003	0.00017	0.99350	0.99415

You can learn more about the R code for this example at the online version of this book, [Actuarial Community \(2025\)](#).

---

**Discussion.** One of the many interesting things about the leave-one-out special case is the ability to replicate estimates exactly. That is, when the size of the fold is only one, then there is no additional uncertainty induced by the cross-validation. This means that analysts can exactly replicate work of one another, an important consideration.

Jackknife statistics were developed to understand precision of estimators, producing estimators of bias and standard deviation in equations (8.3) and (8.4). This crosses into goals that we have associated with bootstrap techniques, not cross-validation methods. This demonstrates how statistical techniques can be used to achieve different goals.

#### 8.4.3 Cross-Validation and Bootstrap

The bootstrap is useful in providing estimators of the precision, or variability, of statistics. It can also be useful for model validation. The bootstrap approach to model validation is similar to the leave-one-out and  $k$ -fold validation procedures:

- Create a bootstrap sample by re-sampling (with replacement)  $n$  indices in

$\{1, \dots, n\}$ . That will be our *training sample*. Estimate the model under consideration based on this sample.

- The *test*, or *validation sample*, consists of those observations not selected for training. Evaluate the fitted model (based on the training data) using the test data.

Repeat this process many (say  $B$ ) times. Take an average over the results and choose the model based on the average evaluation statistic.

**Example 8.4.4. Wisconsin Property Fund.** Return to [Example 8.3.1](#) where we investigate the fit of the gamma and Pareto distributions on the property fund data. We again compare the predictive performance using the Kolmogorov-Smirnov ( $KS$ ) statistic but this time using the bootstrap procedure to split the data between training and testing samples. The following provides illustrative code.

We did the sampling using  $B = 100$  replications. The average  $KS$  statistic for the Pareto distribution was 0.058 compared to the average for the gamma distribution, 0.262. This is consistent with earlier results and provides another piece of evidence that the Pareto is a better model for these data than the gamma.

You can learn more about the R code for this example at the online version of this book, [Actuarial Community \(2025\)](#).

---

## 8.5 Further Resources and Contributors

Section [8.4.2](#) presented the jackknife statistic as an application of (leave one out) cross-validation methods. Another way to present this material is to consider the historical development. [Efron \(1982\)](#) attributes the jackknife idea to [Quenouille \(1949\)](#). Even in this simpler time before modern computing power became widely available, the jackknife provided a handy tool to estimate the bias and standard deviation for virtually any statistic. In addition, this provided motivation for the 1979 introduction of the bootstrap in [Efron \(1979\)](#) (see also [Efron \(1992\)](#)). The bootstrap provided a tool to understand the uncertainty of a statistic, including the standard deviation.

The presentation in this book, outlined in Chapter [2](#), follows strategies adopted by analysts. We think of the jackknife and the bootstrap as tools that helps one understand qualities of a statistic of interest. In addition, cross-validation is a resampling strategy primarily devoted to model validation. As noted in [Efron \(1982\)](#), the historical development of cross-validation is a bit murkier.

It is a method borne from the very simple strategy of splitting a sample in half, then using a model trained on one half to predict performance in the other half. Comparing cross-validation methods to the jackknifing and bootstrapping techniques, all are based on resampling. In addition, questions of statistical inference naturally overlap with model validation issues, so there is a natural overlap among these methods.

- For further reading, a classic, and still very readable, introduction to the jackknife and bootstrap is provided by [Efron \(1982\)](#).
- Here are some links to learn more about [reproducibility and randomness](#) and how to go [from a random generator to a sample function](#).

#### Contributors

- **Arthur Charpentier**, Université du Québec à Montréal, and **Edward (Jed) Frees**, University of Wisconsin-Madison, are the principal authors of the initial version of this chapter.
  - Chapter reviewers include Yvonne Chueh and Brian Hartman.
- **Edward (Jed) Frees**, University of Wisconsin-Madison and Australian National University, is the author of the second edition of this chapter. Email: [jfrees@bus.wisc.edu](mailto:jfrees@bus.wisc.edu) for chapter comments and suggested improvements.
  - This chapter has benefited significantly from suggestions by Hirokazu (Iwahiro) Iwasawa.

# 9

---

## *Bayesian Statistics and Modeling*

---

*Chapter Preview.* Up to this point in the book, we have focused almost exclusively on the frequentist approach to estimate our various loss distribution parameters. In this chapter, we switch gears and discuss a different paradigm: Bayesianism. These approaches are different as Bayesian and frequentist statisticians disagree on the source of the uncertainty: Bayesian statistics assumes that the observed data sample is fixed and that model parameters are random, whereas frequentism considers the opposite (i.e., the sample data are random, and the model parameters are fixed but unknown).

In this chapter, we introduce Bayesian statistics and modeling with a particular focus on loss data analytics. We begin in Section 9.1 by explaining the basics of Bayesian statistics: we compare it to frequentism and provide some historical context for the paradigm. We also introduce the seminal Bayes' rule that serves as a key component in Bayesian statistics. Then, building on this, we present the main ingredients of Bayesian statistics in Section 9.2: the posterior distribution, the likelihood function, and the prior distribution. Section 9.3 provides some examples of simple cases where the prior distribution is chosen for algebraic convenience, giving rise to a closed-form expression for the posterior; these are called conjugate families in the literature. Section 9.4 is dedicated to cases where we cannot get closed-form expressions and for which numerical integration is needed. Specifically, we discuss two influential Markov chain Monte Carlo samplers: the Gibbs sampler and the Metropolis–Hastings algorithm. We also discuss how to interpret the chains obtained by these methods (i.e., Markov chain diagnostics). Finally, the last section of this chapter, Section 9.5, explains the main computing resources available and gives an illustration in the context of loss data.

## 9.1 A Gentle Introduction to Bayesian Statistics

In Section 9.1, you learn how to:

- Describe qualitatively Bayesianism as an alternative to the frequentist approach.
- Give the historical context for Bayesian statistics.
- Use Bayes' rule to find conditional probabilities.
- Understand the basics of Bayesian statistics.

### 9.1.1 Bayesian versus Frequentist Statistics

Classic frequentist statistics rely on frequentist probability—an interpretation of probability in which an event's probability is defined as the limit of its relative frequency (or propensity) in many, repeatable trials. It draws conclusion from a sample that is one of many hypothetical datasets that could have been collected; the uncertainty is therefore due to the sampling error associated with the sample, while model parameters and various quantities of interest are fixed (but unknown to the experimenter).

**Example 9.1.1. Coin Toss.** Considering the simple case of coin tossing, if we flip a fair coin many times, we expect to see heads about 50% of the time. If we flip the coin only a few times, however, we could see a different sample just by chance. Indeed, there is a non-zero probability of observing all heads (and this even if the sample is very large). Figure 9.1 illustrates this the number of heads observed in 100 samples of five iid tosses; in this specific example, we observe six samples for which all tosses are heads.<sup>1</sup>

Yet, as the sample size increases, the relative frequency of heads should get closer to 50% if the coin is fair. Figure 9.2 reports that, if the number of tosses increases, then relative frequency of heads gets closer to 0.5—the probability of seeing heads on a given coin toss. In other words, increasing the sample size makes the resulting parameter estimate less uncertain, and the experimenter should be reaching a probability of 0.5 in the limit, assuming they can reproduce the experiment an infinite number of times.

<sup>1</sup>Each coin toss can be seen as a Bernoulli random variable, meaning that their sum is a binomial with parameters  $q = 0.5$  and  $m = 5$ . See Chapter ?? for more details.

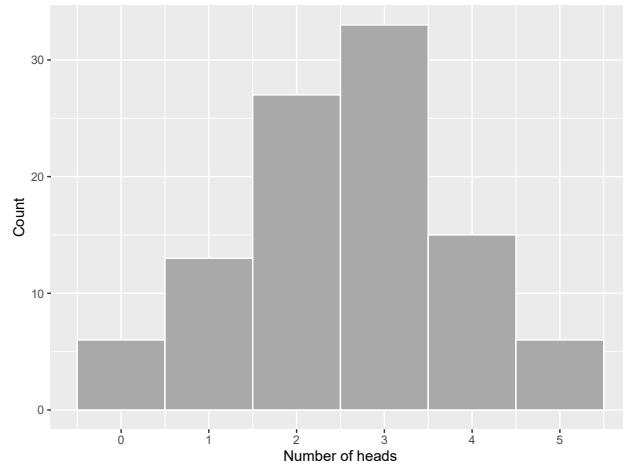


FIGURE 9.1: Frequency histogram of the number of heads in a sample of five data points

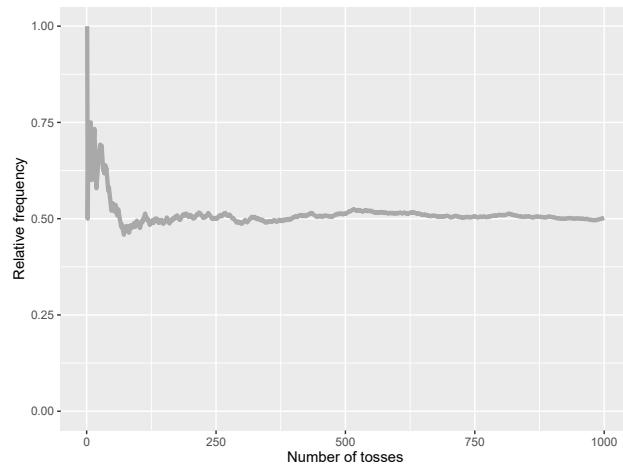


FIGURE 9.2: Cumulative relative frequencies of heads for an increasing sample size

---

Bayesians see things differently: they interpret probabilities as degrees of certainty about some quantity of interest. To find such probabilities, they draw on prior knowledge about those quantities, expressing one's beliefs before some data are taken into account. Then, as data are collected, knowledge about the world is updated, allowing us to incorporate such new information in a consistent manner; the resulting distribution is referred to as the posterior, which summarizes the information in both the prior and the data.

In the context of Bayesian statistics and modeling, this interpretation of probability implies that model parameters are assumed to be random variables—unlike the frequentist approach that considers them fixed. Starting from the prior distribution, the data—summarized via the likelihood function—are used to update the prior distribution and create a posterior distribution of the parameters (see Section 9.2 for more details on the posterior distribution, the likelihood function, and the prior distribution). The influence of the prior distribution on the posterior distribution becomes weaker as the size of the observed data sample increases: the prior information is less and less relevant as new information comes in.

---

**Example 9.1.1. Coin Toss, continued.** We now reconsider the coin tossing experiment above through a Bayesian lens. Let us first assume that we have a (potentially unfair) coin, and we wish to understand the probability of obtaining heads, denoted by  $q$  in this example. Consistent with the Bayesian paradigm, this parameter is random; let us assume that the random variable associated with the probability of observing heads is denoted by  $Q$ . For simplicity, we assume that we do not have prior information on the specific coin under investigation.<sup>2</sup> Assuming again that our sample contains only five iid tosses, we know that the probability of observing  $x$  heads is given by the binomial distribution with  $m = 5$  such that

$$p_{X|Q=q}(x) = \Pr(X = x | Q = q) = \binom{5}{x} q^x (1 - q)^{5-x}, \quad x \in \{0, 1, \dots, 5\},$$

where  $0 \leq q \leq 1$ , which emphasizes the fact that this probability depends on parameter  $q$  by explicitly conditioning on it (unlike the notation used so far in this book, note that we append subscripts to the various pdf and pmf in this chapter to denote the random variables under study; this additional notation allows us to consider the pdf and pmf of different random variables in the same problem).

---

<sup>2</sup>Specifically, we use a uniform over  $[0, 1]$  for our prior distribution. As explained in Section 9.2.3, this type of prior is said to be noninformative.

Let us generate a sample of these five tosses:

```
set.seed(1)
nbheads <- c(1)
num_flips <- 5
coin <- c("heads", "tails")
flips <- sample(coin, size = 5, replace = TRUE)
nbheads <- sum(flips == "heads")
cat("Number of heads:", nbheads)
```

Number of heads: 3

Based on this simulation, we obtain a data sample that contains three heads and two tails. Therefore, using Bayesian statistics, we can show that

$$f_{Q|X=3}(q) \propto q^3(1-q)^2,$$

where  $\propto$  means proportional to (note that obtaining this equation requires some tools that will be introduced in Section 9.2).<sup>3</sup> Figure 9.3 illustrates this pdf and reports the uncertainty about parameter  $q$  based on this sample of five data points. In this example, one can see that the uncertainty is quite large; this is a by-product of using only five data points. Indeed, based on these five observations, one could argue that the probability should be close to  $\frac{3}{5} = 0.6$ . This Bayesian analysis shows that 0.6 is likely, but that it is also very uncertain—a conclusion that is not direct in the frequentist approach.

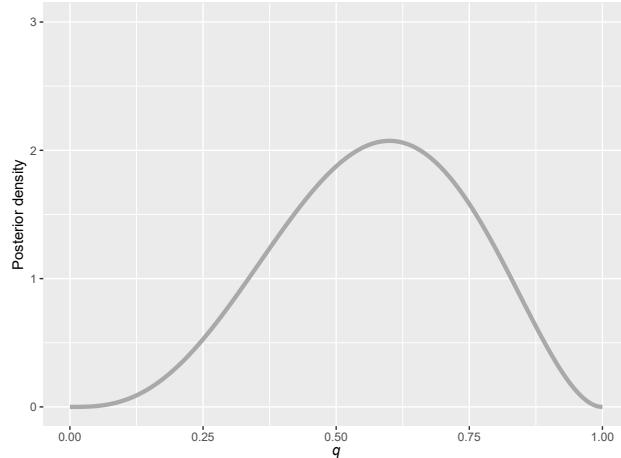


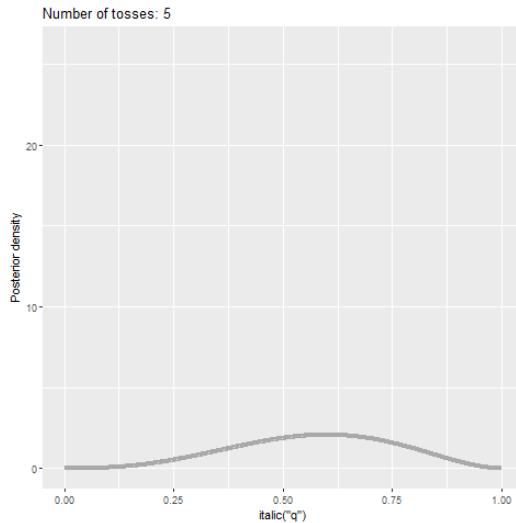
FIGURE 9.3: Posterior probability density function of parameter  $q$  for a sample of five data points

Figure 9.4 reports the analog of Figure 9.2 through a Bayesian lens: we see the

---

<sup>3</sup>This is also an application of the beta-binomial conjugate family that will be explained in Section 9.3.1

evolution of the posterior density of parameter  $q$  as a function of the sample size for the same sample used in Figure 9.2. As we obtain more evidence, the posterior density becomes more concentrated around 0.5—a consequence of using a fair coin in the simulations above. Yet, even if the sample size is 1,000, we still see some parameter uncertainty.



**FIGURE 9.4: Posterior probability density function of parameter  $q$  as a function of the sample size**

**But why be Bayesian?** There are indeed several advantages to the Bayesian approach. First, this approach allows us to describe the entire distribution of parameters conditional on the data and to provide probability statements regarding the parameters that could be interpreted as such. Second, it provides a unified approach for estimating parameters. Some non-Bayesian methods, such as least squares, require a separate approach to estimate variance components. In contrast, in Bayesian methods, all parameters can be treated in a similar fashion. Third, it allows experimenters to blend prior information from other sources with the data in a coherent manner.<sup>4</sup>

**Are there any disadvantages to being Bayesian?** Well, of course: while the Bayesian approach has many advantages, it is not without its disadvan-

<sup>4</sup>There is also a rich history blending prior information with data in loss modeling and in actuarial science, generally speaking; it is known as credibility. In technical terms, credibility theory's main challenge lies in identifying the optimal linear approximation to the mean of the Bayesian predictive density. This is the reason credibility theory shares numerous outcomes with both linear filtering and the broader field of Bayesian statistics. For more details on experience rating using credibility theory, see Chapter ??.

tages. First, it tends to be very computationally demanding (i.e., Bayesian methods often require complex computations, especially when dealing with high-dimensional problems or large datasets). For instance, complex models may not have closed-form solutions and require specialized computational techniques, which can be time-consuming. Second, there is some subjectivity in selecting priors—our initial beliefs and knowledge about the parameters—and this can lead to different results in the end. Third, Bayesian analysis often produces results that can be challenging to communicate effectively to non-experts.

Despite these disadvantages, the Bayesian approach remains powerful and flexible for many actuarial problems.

**Do I need to be a Bayesian to embrace Bayesian statistics?** No, this can be decided on a case-by-case basis. Consider a Bayesian study when you have prior knowledge or beliefs about the parameters, need to explicitly quantify uncertainty in your estimates, have limited data, require a flexible framework for complex models, or when decision-making under uncertainty is a key aspect of your analysis.

Even if one does not want to be a Bayesian truly, they can still recognize the usefulness of some of the methods. Indeed, some modern statistical tools in artificial intelligence and machine learning rely heavily on Bayesian techniques (e.g., Bayesian neural networks, Gaussian processes, and Bayesian classifiers, to name a few).

### 9.1.2 A Brief History Lesson

Interestingly, some have argued that the birth of Bayesian statistics is intimately related to insurance; see, for instance, [Cowles \(2013\)](#). Specifically, the Great Fire of London in 1666—destroying more than 10,000 homes and about 100 churches—led to the rise of insurance as we know it today. Shortly after, the first full-fledged fire insurance company came into existence in England during the 1680s. By the turn of the century, the idea of insurance was well ingrained and its use was booming in England; see, for instance, [Haueter \(2017\)](#). Yet, the lack of statistical models and methods—much needed to understand risk—drove some insurers to bankruptcy.

Thomas Bayes, an English statistician, philosopher and Presbyterian minister, applied his mind to some of these important statistical questions raised by insurers. This culminated into Bayes’ theory of probability in his seminal essay entitled *Essay towards solving a problem in the doctrine of chances*, published posthumously in 1763. This essay laid out the foundation of what we now know as Bayesian statistics.



FIGURE 9.5: Portrait of an unknown Presbyterian clergyman identified as Thomas Bayes in [O'Donnell \(1936\)](#)

Thomas Bayes' work also helped Pierre-Simon Laplace, a famous French scholar and polymath, to develop and popularize the Bayesian interpretation of probability in the late 1700s and early 1800s. He also moved beyond Bayes' essay and generalized his framework. Laplace's efforts were followed by many, and Bayesian thinking continued to progress throughout the years with the help of statisticians like Bruno de Finetti, Harold Jeffreys, Dennis Lindley, and Leonard Jimmie Savage.

Nowadays, Bayesian statistics and modeling is widely used in science, thanks to the increase in computational power over the past 30 years. Actuarial science and loss modeling, more specifically, have also been breeding grounds for Bayesian methodology. So, Bayesian statistics circles back to insurance, in a sense, where it all started.

### 9.1.3 Bayes' Rule

This subsection introduces how the Bayes' rule is applied to calculating conditional probabilities for events.

**Conditional Probability.** The concept of conditional probability considers the relationship between probabilities of two (or more) events happening. In its most simple form, being interested in conditional probability boils down to answering this question: *given that event B happened, how does this affect the probability that A happens?* To answer this question, we can define formally the concept of conditional probability:

$$\Pr(A | B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

To be properly defined, we must assume that  $\Pr(B)$  is larger than zero; that is, event  $B$  is not impossible. Simply put, a conditional probability turns  $B$

into the new probability space, and then cares only about the part of  $A$  that is inside  $B$  (i.e.,  $A \cap B$ ).

**Example 9.1.2. Actuarial Exam Question.** An insurance company estimates that 40% of policyholders who have an extended health policy but no long-term disability policy will renew next year, and 70% of policyholders who have a long-term disability policy but no extended health policy will renew next year. The company also estimates that 50% of their clients who have both policies will renew at least one next year. The company records report that 65% of clients have an extended health policy, 40% have a long-term disability policy, and 10% have both. Using the data above, calculate the percentage of policyholders that will renew at least one policy next year.<sup>5</sup>

**Example Solution.** Let  $E$  be the event that a policyholder has an extended health policy,  $D$  be the event that a policyholder has a long-term disability policy, and  $R$  be the event that a policyholder renews a policy. We are given:

$$\begin{aligned} - \Pr(E) &= 0.65, & - \Pr(D) &= 0.40, & - \Pr(E \cap D) &= 0.10, & - \Pr(R | E \cap D^c) &= 0.40, \\ - \Pr(R | E^c \cap D) &= 0.70, & - \Pr(R | E \cap D) &= 0.50. \end{aligned}$$

We are looking for  $\Pr(R)$ .

Note that

$$\Pr(E \cap D^c) = \Pr(E) - \Pr(E \cap D) = 0.65 - 0.10 = 0.55,$$

and

$$\Pr(E^c \cap D) = \Pr(D) - \Pr(E \cap D) = 0.40 - 0.10 = 0.30.$$

Moreover, note that  $E \cap D^c$ ,  $E^c \cap D$ , and  $E \cap D$  are mutually disjoint, and that

$$\begin{aligned} \Pr(R) &= \Pr(R \cap (E \cap D^c)) + \Pr(R \cap (E^c \cap D)) + \Pr(R \cap (E \cap D)) \\ &= \Pr(R | (E \cap D^c)) \Pr(E \cap D^c) + \Pr(R | (E^c \cap D)) \Pr(E^c \cap D) \\ &\quad + \Pr(R | (E \cap D)) \Pr(E \cap D) \\ &= 0.40 \times 0.55 + 0.70 \times 0.30 + 0.50 \times 0.10 \\ &= 0.48. \end{aligned}$$

**Independence.** If two events are unrelated to one another, we say that they are independent. Specifically,  $A$  and  $B$  are independent if

$$\Pr(A \cap B) = \Pr(A) \Pr(B).$$

---

<sup>5</sup>This question was adapted from the Be An Actuary website. See [here](#) for more details.

For positive probability events, independence between  $A$  and  $B$  is also equivalent to

$$\Pr(A | B) = \Pr(A) \quad \text{and} \quad \Pr(B | A) = \Pr(B),$$

which means that the occurrence of event  $B$  does not have an impact on the occurrence of  $A$ , and vice versa.

**Bayes' Rule.** Intuitively speaking, Bayes' rule provides a mechanism to put our Bayesian thinking into practice. It allows us to update our information by combining the data—from the likelihood—and the prior together to obtain a posterior probability.

**Proposition 9.1.1. Bayes' Rule for Events.** For events  $A$  and  $B$ , the posterior probability of event  $A$  given  $B$  follows

$$\Pr(A | B) = \frac{\Pr(B | A) \Pr(A)}{\Pr(B)},$$

where the law of total probability allows us to find

$$\Pr(B) = \Pr(A) \Pr(B | A) + \Pr(A^c) \Pr(B | A^c).$$

Note, again, that this works as long as event  $B$  is possible (i.e.,  $\Pr(B) > 0$ ).<sup>6</sup>

**Proof.** Bayes' rule may be derived from the definition of conditional probability shown above:

$$\Pr(A | B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

if  $\Pr(B) > 0$ . Similarly,

$$\Pr(B | A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

if  $\Pr(A) > 0$ . Solving for  $\Pr(A \cap B)$  in the last equation and substituting

---

<sup>6</sup>The law of total probability states that the total probability of an event  $B$  is equal to the sum of the probabilities of  $B$  occurring under different conditions, weighted by the probabilities of those conditions. In the case where there are only two different conditions (let us say  $A$  and  $A^c$ ), we simply need to consider these two conditions. In all generality, however, we would need to consider more possibilities if the sample space cannot be divided into only two events.

into the first one yields Bayes' rule:

$$\Pr(A | B) = \frac{\Pr(B | A) \Pr(A)}{\Pr(B)}.$$

Simply put, the posterior probability of event  $A$  given  $B$  is obtained by combining the likelihood of  $B$  given a fixed  $A$ —proxied by  $\Pr(B | A)$ —with the prior probability of observing  $A$ , and then dividing it by the marginal probability of event  $B$  to make sure that the probabilities sum up to one.

**Example 9.1.3. Actuarial Exam Question.** An automobile insurance company insures drivers of all ages. An actuary compiled the probability of having an accident for some age bands as well as an estimate of the portion of the company's insured drivers in each age band:

Age of Driver	Probability of Accident	Portion of Company's Insured Drivers
16-20	0.06	0.08
21-30	0.03	0.15
31-65	0.02	0.49
66-99	0.04	0.28

A randomly selected driver that the company insures has an accident. Calculate the probability that the driver was age 16-20.<sup>7</sup>

**Example Solution.** Let  $B$  be the event of an insured driver having an accident, and let

- $A_1$  be the event related to the driver's age being in the range 16-20,
- $A_2$  be the event related to the driver's age being in the range 21-30,
- $A_3$  be the event related to the driver's age being in the range 31-65,
- $A_4$  be the event related to the driver's age being in the range 66-99.

<sup>7</sup>This question was taken from the Society of Actuaries Sample Questions for Exam P. See [here](#) for more details.

Then,

$$\begin{aligned}\Pr(A_1 | B) &= \frac{\Pr(B | A_1) \Pr(A_1)}{\Pr(B | A_1) \Pr(A_1) + \Pr(B | A_2) \Pr(A_2) + \Pr(B | A_3) \Pr(A_3) + \Pr(B | A_4) \Pr(A_4)} \\ &= \frac{0.06 \times 0.08}{0.06 \times 0.08 + 0.03 \times 0.15 + 0.02 \times 0.49 + 0.04 \times 0.28} \\ &= 0.1584.\end{aligned}$$

#### 9.1.4 An Introductory Example of Bayes' Rule

The example above illustrates how to use Bayes' rule in an academic context; the focus of this book is, nonetheless, data analytics. We therefore also wish to illustrate Bayes' rule by using *real* data. In this introductory example, we use the Singapore auto data `sgautonb` of the R package `CASdatasets` that was already used in Chapter 3.

```
library("CASdatasets")
data(sgautonb)
```

This dataset contains information about the number of car accidents and some risk factors (i.e., the type of the vehicle insured, the age of the vehicle, the sex of the policyholder, and the age of the policyholder grouped into seven categories).<sup>8</sup>

---

**Example 9.1.4. Singapore Insurance Data.** A new insurance company—targeting an older segment of the population—estimates that 20% of their policyholders will be 65 years old and older. The actuaries working at the insurance company believes that the Singapore insurance dataset is credible to understand the accident occurrence of the new company. Based on this information, find the probability that a randomly selected driver who has (at least) one accident, is 65 years or older.

**Example Solution.** Let  $O$  denote the event related to the policyholder being 65 years old and older (i.e., Age Category 6 in the dataset), and  $A$  the event of a policyholder having at least an accident. Using Bayes' rule, we have that

---

<sup>8</sup>The data are from the General Insurance Association of Singapore, an organization consisting of non-life insurers in Singapore. These data contain the number of car accidents for  $n = 7,483$  auto insurance policies with several categorical explanatory variables and the exposure for each policy.

$$\Pr(O | A) = \frac{\Pr(A | O) \Pr(O)}{\Pr(A)},$$

where the prior probability  $\Pr(O)$  is given by the problem statement:  $\Pr(O) = 0.20$ . This implies that  $\Pr(O^c) = 1 - 0.20 = 0.80$ . From the Singapore insurance data, we know that  $\Pr(A | O) = 0.1082803$  and  $\Pr(A | O^c) = 0.06415506$ , which allow us to use the law of total probability to obtain:

$$\Pr(A) = \Pr(A | O) \Pr(O) + \Pr(A | O^c) \Pr(O^c).$$

```
# Example 9.1.4 Illustrative Code
n <- length(sgautonb$AgeCat)
n0 <- sum(sgautonb$AgeCat == 6)
n0c <- sum(sgautonb$AgeCat != 6)
nAand0 <- sum(sgautonb$AgeCat == 6 & sgautonb$Clm_Count > 0)
nAand0c <- sum(sgautonb$AgeCat != 6 & sgautonb$Clm_Count > 0)

PA0 <- nAand0/n0
PA0c <- nAand0c/n0c

POA <- PA0 * 0.2/(PA0 * 0.2 + PA0c * 0.8)
cat("The probability that policyholder having accident \n is 65 years old and older is",
    POA)
```

The probability that policyholder having accident  
is 65 years old and older is 0.296739115

The probability that a randomly selected driver who has (at least) one accident, is 65 years or older is therefore about 29.7

---

In the next section, we expand on the idea of Bayes' rule and apply it to slightly more general cases involving random variables instead of events.

---

## 9.2 Building Blocks of Bayesian Statistics

In Section 9.2, you learn how to:

- Describe the main components of Bayesian statistics; that is, the posterior distribution, the likelihood function, and the prior distribution.
- Summarize the different classes of priors used in practice.

Proposition 9.1.1 above deals with the elementary case of Bayes' rule for events. Although this version of Bayes' rule is useful to understand the foundation of Bayesian statistics, we will need slightly more general versions of it to achieve our aim. Specifically, Proposition 9.1.1 needs to be generalized to the case of random variables.

Let us first consider the case of discrete random variables. Assume  $X$  and  $Y$  are both discrete random variables that allow for the following joint pmf of

$$p_{X,Y}(x,y) = \Pr(X = x \text{ and } Y = y)$$

as well as the following marginal distributions for  $X$  and  $Y$ :

$$p_X(x) = \Pr(X = x) = \sum_k p_{X,Y}(x,k) \quad \text{and} \quad p_Y(y) = \Pr(Y = y) = \sum_k p_{X,Y}(k,y),$$

respectively. Using the result of Proposition 9.1.1 and setting event  $A$  as  $\{Y = y\}$  and  $B$  as  $\{X = x\}$  yields

$$p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x) p_Y(y)}{p_X(x)},$$

where  $p_{Y|X=x}(y) = \Pr(Y = y | X = x)$  is the conditional pmf of  $Y$  conditional on  $X$  being equal to  $x$ . Using the law of total probability,

$$p_X(x) = \sum_k p_{X,Y}(x,k) = \sum_k p_{X|Y=k}(x) p_Y(k),$$

we can rewrite the denominator above to get the following version of Bayes' rule:

$$p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x) p_Y(y)}{\sum_k p_{X|Y=k}(x) p_Y(k)}.$$

We can also obtain a similar Bayes' rule for continuous random variables by replacing probability mass functions by probability density functions, and sums by integrals.

**Proposition 9.2.1. Bayes' Rule for Continuous Random Variables.** For two continuous random variables  $X$  and  $Y$ , the conditional probability density function of  $Y$  given  $X = x$  follows

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x) f_Y(y)}{f_X(x)},$$

where the marginal distributions of  $X$  and  $Y$  are given as follows:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,u) du \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(u,y) du,$$

respectively. Similar to the discrete random variable case, we can swap the denominator of the equation above for

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,u) du = \int_{-\infty}^{\infty} f_{X|Y=u}(x) f_Y(u) du$$

by using the law of total probability.

**Proof.** Bayes' rule for continuous random variables may be derived from the definition of conditional probability density functions:

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

if  $f_X(x) > 0$ . Similarly,

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

if  $f_Y(y) > 0$ . Solving for  $f_{X,Y}(x,y)$  in the last equation and substituting into the first one yields Bayes' rule for continuous random variables:

$$f_{Y|X=x}(y) = \frac{f_{X|Y=y}(x) f_Y(y)}{f_X(x)}.$$

---

Note that one can mix the discrete and continuous definitions of Bayes' rule to accommodate for cases where the parameters have continuous random variables and the observations are expressed via discrete random variables, or vice versa.

### 9.2.1 Posterior Distribution

Model parameters are assumed to be random variables under the Bayesian paradigm, meaning that Bayes' rule for (discrete or continuous) random variables can be applied to update the prior knowledge about parameters by using new data. This is indeed similar to the process used in Section 9.1.1.

Let us consider only one unknown model parameter  $\theta$  associated with random

variable  $\Theta$  for now.<sup>9</sup> Further, consider  $n$  observations

$$\mathbf{x} = (x_1, x_2, \dots, x_n),$$

which are realizations of the collection of random variables

$$\mathbf{X} = (X_1, X_2, \dots, X_n).$$

If  $Y$  in Proposition 9.2.1 is replaced by  $\Theta$  and  $X$  by  $\mathbf{X}$ , we obtain

$$f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) = \frac{f_{\mathbf{X}|\Theta=\theta}(\mathbf{x}) f_{\Theta}(\theta)}{f_{\mathbf{X}}(\mathbf{x})},$$

which represents the posterior distribution of the model parameter after updating the distribution based on the new observations  $\mathbf{x}$ , and where

- $f_{\mathbf{X}|\Theta=\theta}(\mathbf{x})$  is the likelihood function, also known as the conditional joint pdf of the observations assuming a given value of parameter  $\theta$ ,
- $f_{\Theta}(\theta)$  is the unconditional pdf of the parameter that represents the prior information, and
- $f_{\mathbf{X}}(\mathbf{x})$  is the marginal likelihood, which is a constant term with respect to  $\theta$ , making the posterior density integrate to one.

In other words, Bayes' rule provides a way to update the prior distribution of the parameter into a posterior distribution—by considering the observations  $\mathbf{x}$ .

Note that the marginal likelihood is constant once we have the observations. It does not depend on  $\theta$  and does not impact the overall shape of the pdf: it only provides the adequate scaling to ensure that the density integrates to one. For this reason, it is common to write down the posterior distribution using a proportional relationship instead:

$$f_{\Theta|\mathbf{X}=\mathbf{x}}(\theta) \propto \underbrace{f_{\mathbf{X}|\Theta=\theta}(\mathbf{x})}_{\text{Likelihood}} \underbrace{f_{\Theta}(\theta)}_{\text{Prior}}.$$

**Example 9.2.1. A Problem Inspired from Meyers (1994).** A car insurance pays the following (independent) claim amounts on an automobile insurance policy:

$$1050, \quad 1250, \quad 1550, \quad 2600, \quad 5350, \quad 10200.$$

---

<sup>9</sup>For the sake of simplicity, we only consider one parameter in our derivation here. Note that, later, we will consider cases with more than one parameter and that this extension does not change the bulk of our results and derivations.

The amount of a single payment is distributed as a single-parameter Pareto distribution with  $\theta = 1000$  and  $\alpha$  unknown, such that

$$f_{X_i|A=\alpha}(x_i) = \frac{\alpha 1000^\alpha}{x_i^{\alpha+1}}, \quad x_i \in \mathbb{R}_+.$$

We assume that the prior distribution of  $\alpha$  is given by a gamma distribution with shape parameter 2 and scale parameter 1, and its pdf is given by

$$f_A(\alpha) = \alpha e^{-\alpha}, \quad \alpha \in \mathbb{R}_+.$$

Find the posterior distribution of parameter  $\alpha$ .

**Example Solution.** The likelihood function is constructed by multiplying the pdf of the single payment amounts because they are independent; that is,

$$f_{\mathbf{X}|A=\alpha}(\mathbf{x}) = \prod_{i=1}^6 f_{X_i|A=\alpha}(x_i) = \frac{\alpha^6 1000^{6\alpha}}{\prod_{i=1}^6 x_i^{\alpha+1}} = \alpha^6 e^{-5.66518\alpha - 41.44653}.$$

The posterior distribution is given by

$$f_{A|\mathbf{X}=\mathbf{x}}(\alpha) = \frac{\alpha^7 e^{-6.66518\alpha - 41.44653}}{\int_0^\infty \alpha^7 e^{-6.66518\alpha - 41.44653} d\alpha} = \frac{\alpha^7 e^{-6.66518\alpha}}{\int_0^\infty \alpha^7 e^{-6.66518\alpha} d\alpha}.$$

Interestingly, we do not need to solve the integral in the denominator to find this distribution. As we know that the results should be a proper pdf and that the numerator looks like a gamma distribution, we can deduce that

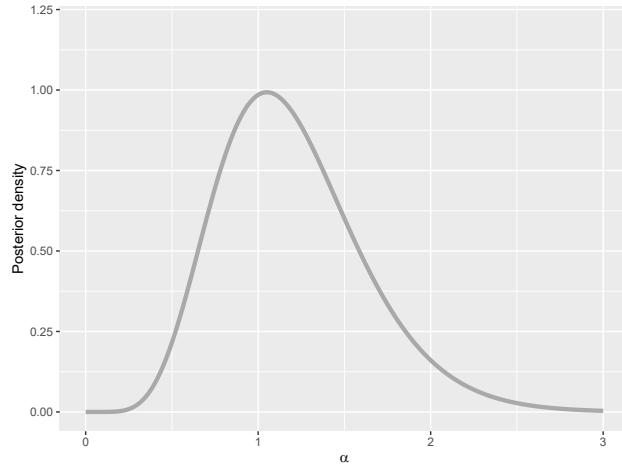
$$f_{A|\mathbf{X}=\mathbf{x}}(\alpha) = \frac{6.66518^8}{\Gamma(8)} \alpha^7 e^{-6.66518\alpha},$$

which is a gamma distribution with shape parameter 8 and scale parameter  $\frac{1}{6.66518}$ . Figure 9.6 reports the posterior distribution of  $\alpha$ .

The discussion above considered continuous random variables, but the same logic can be applied to discrete random variables by replacing probability density functions by probability mass functions.

---

**Example 9.2.2. Coin Toss Revisited.** Assume that you observe three heads out of five (independent) tosses. Each toss has a probability of  $q$  of observing heads and  $1-q$  of observing tails. Find the posterior distribution of  $q$  assuming a uniform prior distribution over the interval  $[0, 1]$ .

FIGURE 9.6: Posterior densities of parameter  $\alpha$ 

**Example Solution.** The prior distribution of  $q$  is given by

$$f_Q(q) = 1, \quad q \in [0, 1].$$

Assuming the likelihood function conditional on  $Q = q$  is given by a binomial distribution with  $m = 5$  and  $x = 3$ ,

$$p_{X|Q=q}(x) = \binom{5}{3} q^3 (1-q)^2,$$

we have that the posterior distribution of  $q$  is given by

$$f_{Q|X=3}(q) \propto p_{X|Q=q}(x) f_Q(q) = q^3 (1-q)^2,$$

which is a beta distribution with  $a = 4$ ,  $b = 3$ , and  $\theta = 1$ ; that is, we can easily deduce that

$$f_{Q|X=3}(q) = \frac{\Gamma(7)}{\Gamma(4)\Gamma(3)} q^3 (1-q)^2.$$

---

In the following subsections, we will discuss at greater length the two main building blocks used to build the posterior distribution: the likelihood function and the prior distribution.

### 9.2.2 Likelihood Function

The likelihood function is a fundamental concept in statistics. It is used to estimate the parameters of a statistical model based on observed data. As mentioned in previous chapters, the likelihood function can be used to find the

maximum likelihood estimator. In Bayesian statistics, the likelihood function is used to update the prior based on the evidence (or data).

As explained above and in Chapter ??, the likelihood function is defined as the conditional joint pdf or pmf of the observed data, given the model parameters. In other words, it is the probability of observing the data given a specific parameter values.

Mathematically, the likelihood function is written as  $f_{\mathbf{X}|\Theta=\theta}(x)$  (for continuous random variables) or  $p_{\mathbf{X}|\Theta=\theta}(x)$  (for discrete random variables). Note that, throughout the book, the notation  $L(\theta|\mathbf{x})$  has also been used for the likelihood function, and we will use both interchangeably in this chapter.

**Special Case: Independent and Identically Distributed Observations.** Oftentimes, in many problems and real-world applications, the observations are assumed to be iid. If they are, then we can easily write the likelihood function as:

$$f_{\mathbf{X}|\Theta=\theta}(\mathbf{x}) = \prod_{i=1}^n f_{X_i|\Theta=\theta}(x_i) \quad \text{or} \quad p_{\mathbf{X}|\Theta=\theta}(\mathbf{x}) = \prod_{i=1}^n p_{X_i|\Theta=\theta}(x_i).$$

### 9.2.3 Prior Distribution

In the Bayesian paradigm, the prior distribution represents our knowledge or beliefs about the unknown parameters before we observe any data. It is a probability distribution that expresses the uncertainty about the values of the parameters. The prior distribution is typically specified by choosing a family of probability distributions and selecting specific values for its parameters.

The choice of prior distribution is subjective and often based on external information or previous studies. In some cases, noninformative priors can be used, which represent minimal prior knowledge or assumptions about the parameters. In other cases, informative and weakly informative priors can be used, which incorporate prior knowledge or assumptions based on external sources. The selection of the prior distribution should be carefully considered, and sensitivity analysis can be performed to assess the robustness of the results to different prior assumptions.

**Why Does It Matter?** The choice of prior distribution can have a significant impact on the results of a Bayesian analysis. Different prior distributions can lead to different posterior distributions, which are the updated probability distributions for the parameters after we observe the data. Therefore, it is important to choose a prior distribution that reflects our prior knowledge or beliefs about the parameters.

### Informative and Weakly Informative Priors

Informative and weakly informative priors are terms used to describe the amount of prior knowledge or beliefs that is incorporated into a statistical model. Informative priors contain substantial prior knowledge about the parameters of a model, while weakly informative priors contain moderate prior knowledge.

Informative priors are useful when there is strong, potentially subjective prior information available about the model parameters, which can help to constrain the posterior distribution and improve inference. For example, in an insurance claims analysis study, an informative prior may be used to incorporate previous knowledge, such as the results of a previous claims study.

On the other hand, weakly informative priors are used when there is some—yet little—prior knowledge available or when the goal is to allow the data to drive the analysis. Weakly informative priors are designed to mildly impact the posterior distribution and are often chosen based on principles such as symmetry or scale invariance.

Overall, the choice of prior depends on the specific problem at hand and the available prior knowledge or beliefs. Informative priors can be useful when prior information is available and can improve the precision of the posterior distribution. In contrast, weakly informative priors can be useful when the goal is to allow the data to drive the analysis and avoid imposing strong prior assumptions.

**Example 9.2.3. Actuarial Exam Question.** You are given:

- Annual claim frequencies follow a Poisson distribution with mean  $\lambda$ .
- The prior distribution of  $\lambda$  has the following pdf:

$$f_{\Lambda}(\lambda) = (0.3)\frac{1}{6}e^{-\frac{\lambda}{6}} + (0.7)\frac{1}{12}e^{-\frac{\lambda}{12}}, \quad \text{where } \lambda > 0.$$

Ten claims are observed for an insured in Year 1. Calculate the expected value of the posterior distribution of  $\lambda$ .<sup>10</sup>

<sup>10</sup>This question is a modified version of Sample Question 184 of the Society of Actuaries Exam C sample questions.

**Example Solution.** The posterior distribution can be found from:

$$\begin{aligned} f_{\Lambda|X=10}(\lambda) &= \frac{p_{X|\Lambda=\lambda}(10)f_{\Lambda}(\lambda)}{p_X(10)} \\ &= \frac{\frac{e^{-\lambda}\lambda^{10}}{10!} \left( (0.3)\frac{1}{6}e^{-\frac{\lambda}{6}} + (0.7)\frac{1}{12}e^{-\frac{\lambda}{12}} \right)}{\int_0^{\infty} \frac{e^{-\lambda}\lambda^{10}}{10!} \left( (0.3)\frac{1}{6}e^{-\frac{\lambda}{6}} + (0.7)\frac{1}{12}e^{-\frac{\lambda}{12}} \right) d\lambda} \\ &= \frac{\lambda^{10} \left( \frac{0.3}{6}e^{-\frac{7\lambda}{6}} + \frac{0.7}{12}e^{-\frac{13\lambda}{12}} \right)}{121050}. \end{aligned}$$

The posterior mean is therefore given by

$$\begin{aligned} E[\Lambda | X = 10] &= \frac{1}{121050} \int_0^{\infty} \lambda^{11} \left( \frac{0.3}{6}e^{-\frac{7\lambda}{6}} + \frac{0.7}{12}e^{-\frac{13\lambda}{12}} \right) d\lambda \\ &= \frac{1}{118170} \left( \frac{0.3}{6}(11!)(6/7)^{12} + \frac{0.7}{12}(11!)(12/13)^{12} \right) \\ &= 9.95442. \end{aligned}$$

### Noninformative Priors

It is possible to take the idea of weakly informative priors to the extreme by using noninformative priors. A noninformative prior is a prior distribution that is intentionally chosen to allow the data to have a more decisive influence on the posterior distribution rather than being overly influenced by prior beliefs or assumptions.

Noninformative priors can take different forms, such as flat priors, for instance. A flat prior assigns equal probability to all possible parameter values without additional information or assumptions.

---

**Example 9.2.4. Informative Versus Noninformative Priors.** You wish to investigate the impact of having informative and noninformative priors on a claim frequency analysis. Assume that the claim frequency for each policy follows a Bernoulli random variable with a probability of  $q$  such that

$$q_{X_i|Q=q}(x_i) = q^{x_i}(1-q)^{1-x_i}, \quad x_i \in \{0, 1\},$$

where  $q \in [0, 1]$ , and consider two different prior distributions:

- Informative: Based on past experience, you know that the claim probability is typically less than 5%, thus justifying the use of a uniform distribution over  $[0, 0.05]$ .

- Noninformative: You do not wish your posterior distribution to be impacted by your prior assumption and simply select a uniform distribution over the domain of  $q$ , which is  $[0, 1]$ .

Using the first 100 lines of the Singapore insurance dataset (see Example 9.1.4 for more details on this dataset), find the two posterior distributions as well as the posterior expected value of the probability  $q$  under both prior assumptions.

**Example Solution.** Let us start with the informative prior, where

$$f_Q(q) = \frac{1}{0.05 - 0} = 20, \quad \text{if } q \in [0, 0.05],$$

and zero otherwise. In this case, assuming  $x = \sum_{i=1}^{100} x_i$ , the posterior density is given by

$$\begin{aligned} f_{Q|\mathbf{x}=\mathbf{x}}(q) &\propto f_{\mathbf{X}|Q=q}(\mathbf{x})f_Q(q) \\ &\propto \prod_{i=1}^{100} q^{x_i} (1-q)^{1-x_i} \\ &= q^x (1-q)^{100-x}, \quad \text{if } 0 \leq q \leq 0.05, \end{aligned}$$

and zero otherwise. We can numerically obtain the shape of this posterior distribution by dividing  $q^x (1-q)^{100-x}$  by

$$\int_0^{0.05} q^x (1-q)^{100-x} dq.$$

Note that this prior makes it impossible for the estimated frequency to be greater than 0.05.

The second prior is still uniform, but over  $[0, 1]$  this time, which is given mathematically by

$$f_Q(q) = \frac{1}{1-0} = 1, \quad \text{if } q \in [0, 1],$$

and zero otherwise, leading to the following posterior distribution:

$$\begin{aligned} f_{Q|\mathbf{x}=\mathbf{x}}(q) &\propto f_{\mathbf{X}|Q=q}(\mathbf{x})f_Q(q) \\ &\propto \prod_{i=1}^{100} q^{x_i} (1-q)^{1-x_i} \\ &= q^x (1-q)^{100-x}, \quad \text{if } 0 \leq q \leq 1, \end{aligned}$$

and zero otherwise.

```

qs <- seq(from = 0, to = 0.12, by = 0.0001)
x <- sum(sgautonb$Clm_Count[1:100])

integrandposterior1 <- function(q) {
  q^x * (1 - q)^(100 - x) * ifelse(q >= 0 & q <= 0.05, 1, 0)
}
marglikelihood1 <- integrate(integrandposterior1, 0, 1, abs.tol = .Machine$double.eps^2)$value
posterior1 <- integrandposterior1(qs)/marglikelihood1

integrandposterior2 <- function(q) {
  q^x * (1 - q)^(100 - x) * ifelse(q >= 0 & q <= 1, 1, 0)
}
marglikelihood2 <- integrate(integrandposterior2, 0, 1, abs.tol = .Machine$double.eps^2)$value
posterior2 <- integrandposterior2(qs)/marglikelihood2

```

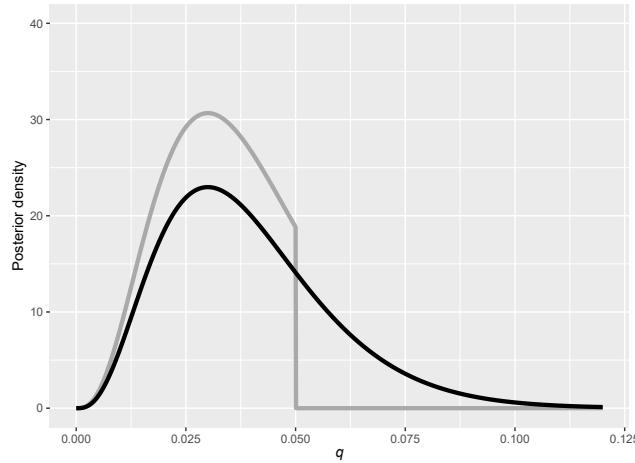


FIGURE 9.7: Posterior densities based on informative (gray) and non-informative priors (black)

We also wish to obtain the expected value of  $q$  for both posterior distribution. This can be obtained by numerically integrating the following equation:

$$\text{E}[Q|\mathbf{X} = \mathbf{x}] = \int_0^1 q f_{Q|\mathbf{X}=\mathbf{x}}(q) dq.$$

```

integrandexpvalue1 <- function(q) {
  integrandposterior1(q)/marglikelihood1 * q
}
expectedvalue1 <- integrate(integrandexpvalue1, 0, 1, abs.tol = .Machine$double.eps^2)$value
cat("The posterior expected value of the parameter \n"
    "when using the informative prior is",
    expectedvalue1)

```

The posterior expected value of the parameter

when using the informative prior is 0.0304525117

```
integrandexpvalue2 <- function(q) {
  integrandposterior2(q)/marglikelihood2 * q
}
expectedvalue2 <- integrate(integrandexpvalue2, 0, 1, abs.tol = .Machine$double.eps^2)$value
cat("The posterior expected value of the parameter \n"
    "when using the noninformative prior is",
    expectedvalue2)
```

The posterior expected value of the parameter

when using the noninformative prior is 0.0392156863

As one can see, these values are different, meaning that the prior distribution can have a material impact on the posterior distribution. One should therefore be careful when selecting a prior distribution.

---

### Improper Priors

An improper prior is a prior distribution that is not a proper probability distribution, meaning that it does not integrate (or sum) to one over the entire parameter space. Improper priors can be used in Bayesian analyses, but they require careful handling because they can lead to improper posterior distributions.

Improper priors are typically used when there is little or no prior information about the parameter of interest—some noninformative priors are indeed improper—and they can be thought of as representing a very diffuse or non-committal prior belief. For instance, the uniform distribution on an infinite interval is a common choice of improper prior.

---

**Example 9.2.5. Improper Prior, Proper Posterior.** Let us assume a random sample  $\mathbf{x}$  of size  $n$ , which is a realization of the collection of random variables  $\mathbf{X} = (X_1, X_2, \dots, X_n)$ . Further, assume that each random variable  $X_i$  is independent and normally distributed with mean of  $\mu$  and variance of 1:

$$f_{X_i|M=\mu}(x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right), \quad x_i \in \mathbb{R},$$

where  $\mu$  is a (random) parameter. Obtain the posterior distribution of  $\mu$  assuming that its prior distribution is improper and given by  $f_M(\mu) \propto 1$ , where  $\mu \in \mathbb{R}$ .

**Example Solution.** According to Bayes' rule, we have that

$$f_{M|\mathbf{X}=\mathbf{x}}(\mu) = \frac{f_{\mathbf{X}|M=\mu}(\mathbf{x}) f_M(\mu)}{f_{\mathbf{X}}(\mathbf{x})} \propto \prod_{i=1}^n f_{X_i|M=\mu}(x_i)$$

because  $f_M(\mu) \propto 1$  and  $f_{\mathbf{X}}(\mathbf{x})$  does not depend on  $\mu$ . Using the equation above, we can obtain the posterior distribution by simplifying the following equation:

$$\begin{aligned} f_{M|\mathbf{X}=\mathbf{x}}(\mu) &\propto \left( \frac{1}{\sqrt{2\pi}} \right)^n \exp \left( -\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 \right) \\ &\propto \exp \left( -\frac{1}{2} \left( \sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \right) \right) \\ &\propto \exp \left( -\frac{n}{2} \left( \frac{\sum_{i=1}^n x_i^2}{n} - \frac{2\mu \sum_{i=1}^n x_i}{n} + \mu^2 \right) \right) \\ &\propto \exp \left( -\frac{n}{2} \left( -\frac{2\mu \sum_{i=1}^n x_i}{n} + \mu^2 \right) \right) \\ &\propto \exp \left( -\frac{n}{2} \left( \mu - \frac{\sum_{i=1}^n x_i}{n} \right)^2 \right) \\ &\propto \frac{1}{\sqrt{2\pi \frac{1}{n}}} \exp \left( -\frac{1}{2} \frac{\left( \mu - \frac{\sum_{i=1}^n x_i}{n} \right)^2}{\frac{1}{n}} \right), \end{aligned}$$

which is a normal distribution with mean  $\frac{\sum_{i=1}^n x_i}{n}$  and variance  $\frac{1}{n}$ . Interestingly, this posterior distribution is proper even though the prior distribution was improper.

---

Special care is needed when dealing with improper priors. Indeed, if one can derive the posterior distribution in closed form and show that it is proper—like in Example 9.2.5—it should not be a concern. On the other hand, in cases where the posterior distribution cannot be obtained in closed form, there is no assurance that the posterior will be proper and extra attention is required.

#### Choice of the Prior Distribution

The selection of a prior in Bayesian statistics is a crucial step that reflects the experimenter's prior beliefs, knowledge, or assumptions about the parameters of interest. There are different approaches to selecting priors.

1. Informative priors are generally based on the experimenter's subjec-

tive beliefs, knowledge, or experience. For instance, one might have a subjective belief that a parameter is likely to fall within a certain range, and this belief is formalized as an informative prior distribution.

2. Noninformative priors are chosen to be minimally informative, expressing little or no prior information about the parameters. For example, uniform priors are commonly used as noninformative priors, expressing a lack of prior preference for any particular parameter value.
3. Empirical Bayes priors rely on the data itself, combining empirical information with Bayesian methodology. This can be done by estimating a prior distribution hyperparameter by using the observed data to inform the prior distribution.
4. Priors that rely on expert elicitation involves seeking input from domain experts to inform the prior. For instance, the experimenter might have additional knowledge about the problem at hand and use a prior distribution that represents their beliefs about the parameters.

### Prior Sensitivity Analysis

Prior sensitivity analysis is an important step in Bayesian modeling processes. It refers to the examination and evaluation of the impact of different prior assumptions on the results of a statistical analysis. In other words, such analyses aim to verify the robustness of the conclusions drawn from Bayesian inference to the choice of the prior distribution. By exploring a range of plausible prior distributions, experimenters can gain insights into how much the choice of prior influences the final results and whether those conclusions remain consistent under different prior assumptions.

For instance, prior distributions may significantly influence the posterior estimates, leading to different conclusions. Some of these might be subjective (i.e., informative priors) or based on expert knowledge, and assessing the impact of such assumptions promotes transparency and objectivity in the analysis.

---

### 9.3 Conjugate Families

---

In Section 9.3, you learn how to:

- Describe three specific classes of conjugate families.

- Use conjugate distributions to determine posterior distributions of parameters.
  - Understand the pros and cons of conjugate family models.
- 

In Bayesian statistics, if a posterior distribution comes from the same distribution as the prior distribution, the prior and posterior are called conjugate distributions. Note that both posterior and prior have similar shapes but will have different parameters, generally speaking.

**But Why?** Two main reasons explain why conjugate families have been so popular historically:

1. They are easy to use from a computational standpoint: posterior distributions in most conjugate families can be obtained in closed form, making this class of models easy to use even if we do not have access to computing power.
  2. They tend to be easy to interpret: posterior distributions are compromises between data and prior distributions. Having both prior and posterior distributions in the same family—but with different parameters—allows us to understand and quantify how the data changed our initial assumptions.
- 

### 9.3.1 The Beta–Binomial Conjugate Family

The first conjugate family that we investigate in this book is the beta–binomial family. Let  $\mathbf{X} = (X_1, X_2, \dots, X_m)$  represent a sample of iid Bernoulli random variables such that

$$X_i = \begin{cases} 1 & \text{if success} \\ 0 & \text{if failure} \end{cases},$$

with probabilities  $q$  and  $1 - q$ , respectively. Let us further define  $x = \sum_{i=1}^m x_i$  the sum of the realized successes.

We know from elementary probability that  $X = \sum_{i=1}^m X_i$  follows a binomial distribution (i.e., the number of successes  $x$  in  $m$  Bernoulli trials) with unknown probability of success  $q$  in  $[0, 1]$ , similar to the coin tossing case of Example 9.1.1, such that the likelihood function is given by

$$p_{\mathbf{X}|Q=q}(\mathbf{x}) = \binom{m}{x} q^x (1-q)^{m-x}, \quad x \in \{0, 1, \dots, m\},$$

where  $x = \sum_{i=1}^m x_i$ . The latter represents our evidence. Then, we combine it

with its usual conjugate prior—the beta distribution with parameters  $a$  and  $b$ . The pdf of the beta distribution is given as follows:

$$f_Q(q) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} q^{a-1} (1-q)^{b-1}, \quad q \in [0, 1],$$

where  $a$  and  $b$  are shape parameters of the beta distribution.<sup>11</sup>

We can now combine the prior distribution—beta—with the likelihood function—binomial—to obtain the posterior distribution.

**Proposition 9.3.1. Beta–Binomial Conjugate Family.** Consider a sample of  $m$  iid Bernoulli experiments  $(X_1, X_2, \dots, X_m)$  each with success probability  $q$ . Further assume that the random variable associated with the success probability,  $Q$ , has a prior that is beta with shape parameters  $a$  and  $b$ . The posterior distribution of  $Q$  is therefore given by

$$f_{Q|\mathbf{X}=\mathbf{x}}(q) = \frac{\Gamma(a+b+m)}{\Gamma(a+x)\Gamma(b+m-x)} q^{a+x-1} (1-q)^{b+m-x-1},$$

where  $x = \sum_{i=1}^m x_i$ , which is a beta distribution with shape parameters  $a+x$  and  $b+m-x$ .

**Proof.** From Section 9.2.1, we know that

$$\begin{aligned} f_{Q|\mathbf{X}=\mathbf{x}}(q) &= \frac{p_{\mathbf{X}|Q=q}(\mathbf{x}) f_Q(q)}{p_{\mathbf{X}}(\mathbf{x})} \propto \binom{m}{x} q^x (1-q)^{m-x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} q^{a-1} (1-q)^{b-1} \\ &\propto q^{a+x-1} (1-q)^{b+m-x-1}. \end{aligned}$$

We therefore only need to find the normalizing constant that ensures that the right-hand of the equation above is a density. Interestingly, the right-hand side looks like a beta distribution; specifically,

$$\begin{aligned} &\int_0^1 q^{a+x-1} (1-q)^{b+m-x-1} dq \\ &= \frac{\Gamma(a+x)\Gamma(b+m-x)}{\Gamma(a+b+m)} \int_0^1 \frac{\Gamma(a+b+m)}{\Gamma(a+x)\Gamma(b+m-x)} q^{a+x-1} (1-q)^{b+m-x-1} dq \\ &= \frac{\Gamma(a+x)\Gamma(b+m-x)}{\Gamma(a+b+m)}, \end{aligned}$$

---

<sup>11</sup>Here, we assume that the domain of the beta is  $[0, 1]$ , meaning that  $\theta = 1$ . For more details, see Chapter ??.

and

$$f_{Q|\mathbf{X}=\mathbf{x}}(q) = \frac{\Gamma(a+b+m)}{\Gamma(a+x)\Gamma(b+m-x)} q^{a+x-1} (1-q)^{b+m-x-1}.$$

**Parameters Versus Hyperparameters.** In this context,  $a$  and  $b$  are called hyperparameters—parameters of the prior. These are different from parameters of the underlying model (i.e.,  $q$  in the beta-binomial family). Hyperparameters are typically assumed and determined by the experimenter, whereas the underlying model parameters are random in the Bayesian context.

**Example 9.3.1. Actuarial Exam Question.** You are given:

- The annual number of claims in Year  $i$  for a policyholder has a binomial distribution with pmf

$$p_{X_i|Q=q}(x_i) = \binom{2}{x} q^{x_i} (1-q)^{2-x_i}, \quad x_i \in \{0, 1, 2\}.$$

- The prior distribution is

$$f_Q(q) = 4q^3, \quad q \in [0, 1].$$

The policyholder had one claim in each of Years 1 and 2. Calculate the Bayesian estimate of the expected number of claims in Year 3.<sup>12</sup>

**Example Solution.** The likelihood function based on this policyholder's number of claims in Years 1 and 2 is given by:

$$p_{\mathbf{X}|Q=q}(\mathbf{x}) = p_{X_1|Q=q}(1) p_{X_2|Q=q}(1) = \binom{2}{1} q^1 (1-q)^1 \binom{2}{1} q^1 (1-q)^1 \propto q^2 (1-q)^2,$$

which is proportional to a binomial pmf with  $m = 4$ , two successes, and a success probability of  $q$ . Because the prior distribution is beta distributed with  $a = 4$  and  $b = 1$ , we know that the posterior distribution of parameter  $q$  is given by

$$\begin{aligned} f_{Q|\mathbf{X}=\mathbf{x}}(q) &= \frac{\Gamma(4+1+4)}{\Gamma(4+2)\Gamma(1+4-2)} q^{4+2-1} (1-q)^{1+4-2-1} \\ &= \frac{\Gamma(9)}{\Gamma(6)\Gamma(3)} q^5 (1-q)^2 \\ &= 168q^5 (1-q)^2, \end{aligned}$$

<sup>12</sup>This question is Sample Question 5 of the Society of Actuaries Exam C sample questions.

which is also a beta distribution with shape parameters 6 and 3, respectively.

The expected number of claim in Year 3 is

$$\mathbb{E}[\mathbb{E}[X_3 | Q = q] | X_1, X_2] = \mathbb{E}[2q | X_1, X_2] = 2\mathbb{E}[q | X_1, X_2],$$

and  $\mathbb{E}[q | X_1, X_2]$  is the expected value of the beta distribution, which is given by

$$\mathbb{E}[q | X_1, X_2] = \frac{6}{6+3} = \frac{2}{3}.$$

Ultimately, this leads to an expected number of claim in Year 3 of  $2\left(\frac{2}{3}\right) = \frac{4}{3}$ .

**Example 9.3.2. Impact of Beta Prior on Posterior.** You wish to investigate the impact of having different beta hyperparameters on the posterior distribution. Assume that the claim frequency for each policy follows a Bernoulli random variable with a probability of  $q$  such that

$$p_{X_i|Q=q}(x_i) = q^{x_i}(1-q)^{1-x_i}, \quad x_i \in \{0, 1\},$$

where  $q \in [0, 1]$ , and consider two different sets of hyperparameters:

- Set 1:  $a = 1$  and  $b = 10$ .
- Set 2:  $a = 2$  and  $b = 2$ .

Figure 9.8 shows the pdf of these two prior distributions. The first prior assumes a small prior mean frequency of  $\frac{1}{11}$ , whereas the second prior distribution has a mean of  $\frac{1}{2}$ .

Using again the first 100 lines of the Singapore insurance dataset (see Example 9.1.4 for more details on this dataset), find the two posterior distributions.

**Example Solution.** The likelihood function associated with the observations is given by

$$p_{\mathbf{X}|Q=q}(\mathbf{x}) = \binom{100}{x} q^x (1-q)^{100-x}, \quad \text{where } x = \sum_{i=1}^{100} x_i,$$

as mentioned already in Example 9.2.4. Combining this likelihood with a beta prior gives a beta posterior:

$$f_{Q|\mathbf{X}=\mathbf{x}}(q) = \frac{\Gamma(a+b+100)}{\Gamma(a+x)\Gamma(b+100-x)} q^{a+x-1} (1-q)^{b+100-x-1},$$

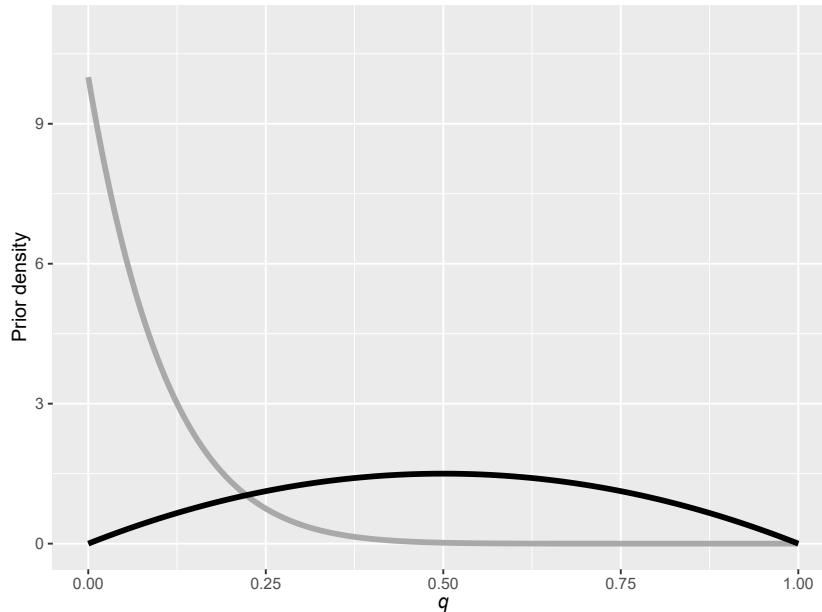


FIGURE 9.8: Beta prior densities:  $a = 1$  and  $b = 10$  (gray), and  $a = 2$  and  $b = 2$  (black)

that can be evaluated for various values of  $a$  and  $b$ . Figure 9.9 reports the two posterior distributions associated with the priors mentioned above.

```
x <- sum(sgautonb$Clm_Count[1:100])

posterior1 <- dbeta(qs, shape1 = 1 + x, shape2 = 10 + 100 - x)
posterior2 <- dbeta(qs, shape1 = 2 + x, shape2 = 2 + 100 - x)

dataposterior <- data.frame(x = qs, y1 = posterior1, y2 = posterior2)

ggplot(dataposterior, aes(x = x, y = y1)) + geom_line(color = "darkgray", lwd = 1.5) +
  geom_line(aes(y = y2), color = "black", lwd = 1.5) + xlim(0, 1) + ylim(0, 35) +
  xlab(expression(italic("q"))) + ylab("Posterior density")
```

---

The prior distribution (and its hyperparameters) clearly have an impact on the posterior distribution. As a general rule of thumb for the beta prior, a higher  $a$  puts more weight on higher values of  $q$  and a higher  $b$  puts more weight on lower values of  $q$ .

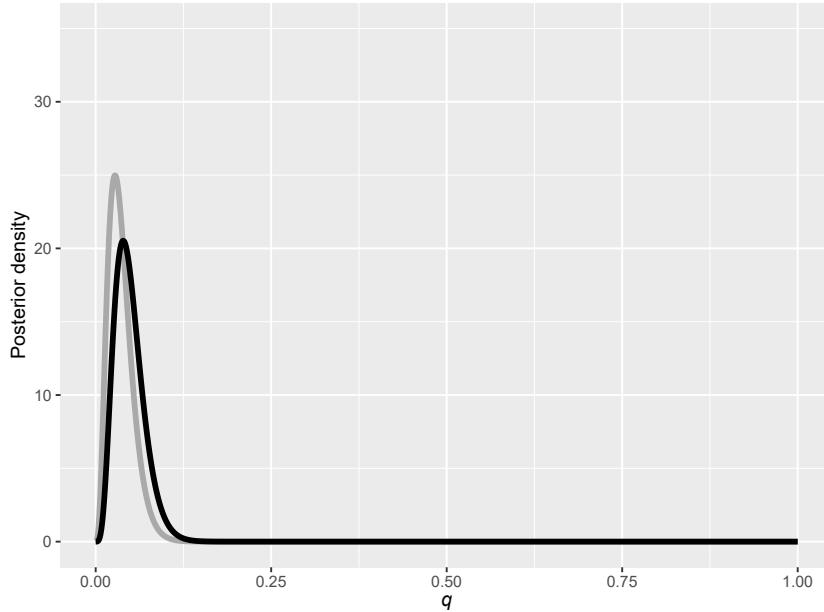


FIGURE 9.9: Posterior densities based on two different priors:  $a = 1$  and  $b = 10$  (gray), and  $a = 2$  and  $b = 2$  (black)

### 9.3.2 The Gamma–Poisson Conjugate Family

We now present a second conjugate family: the gamma–Poisson family. Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  be a sample of iid Poisson random variables such that

$$p_{X_i|\Lambda=\lambda}(x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, \quad x_i \in \mathbb{R}_+.$$

The likelihood function associated with this sample would therefore be given by

$$f_{\mathbf{X}|\Lambda=\lambda}(\mathbf{x}) = \prod_{i=1}^n p_{X_i|\Lambda=\lambda}(x_i) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^x e^{-n\lambda}}{\prod_{i=1}^n x_i!} \propto \lambda^x e^{-n\lambda},$$

where  $x = \sum_{i=1}^n x_i$ . The shape of this likelihood function, as a function of  $\lambda$ , is reminiscent of a gamma distribution, hinting to the fact that this distribution would be a good contender for a conjugate prior. Indeed, if we let the prior distribution be gamma with shape hyperparameter  $\alpha$  and scale hyperparameter  $\theta$ ,

$$f_\Lambda(\lambda) = \frac{1}{\Gamma(\alpha)\theta^\alpha} \lambda^{\alpha-1} e^{-\frac{\lambda}{\theta}}, \quad \lambda \in \mathbb{R}_+,$$

we can show that the posterior distribution of  $\lambda$  is also gamma.

**Proposition 9.3.2. Gamma–Poisson Conjugate Family.** Consider a sample of  $n$  iid Poisson experiments  $(X_1, X_2, \dots, X_n)$ , each with rate parameter  $\lambda$ . Further assume that the random variable associated with the rate,  $\Lambda$ , has a prior that is gamma distributed with shape hyperparameter  $\alpha$  and scale hyperparameter  $\theta$ . The posterior distribution of  $\Lambda$  is therefore given by

$$f_{\Lambda|{\mathbf{X}}={\mathbf{x}}}(\lambda) = \frac{1}{\Gamma(\alpha+x)} \left( \frac{\theta}{n\theta+1} \right)^{\alpha+x} \lambda^{\alpha+x-1} e^{-\frac{\lambda(n\theta+1)}{\theta}},$$

where  $x = \sum_{i=1}^n x_i$ , which is a gamma distribution with shape parameter  $\alpha+x$  and scale parameter  $\frac{\theta}{n\theta+1}$ .

**Proof.** From Section 9.2.1, we know that

$$\begin{aligned} f_{\Lambda|{\mathbf{X}}={\mathbf{x}}}(\lambda) &= \frac{p_{{\mathbf{X}}|\Lambda=\lambda}({\mathbf{x}}) f_\Lambda(\lambda)}{p_{{\mathbf{X}}}({\mathbf{x}})} \propto \lambda^x e^{-n\lambda} \lambda^{\alpha-1} e^{-\frac{\lambda}{\theta}} \\ &\propto \lambda^{\alpha+x-1} e^{-\frac{\lambda(n\theta+1)}{\theta}}, \end{aligned}$$

where  $x = \sum_{i=1}^n x_i$ . We therefore only need to find the normalizing constant that ensures that the right-hand of the equation above is a density. Interestingly, the right-hand side looks like a gamma distribution; specifically,

$$\begin{aligned} &\int_0^\infty \lambda^{\alpha+x-1} e^{-\frac{\lambda(n\theta+1)}{\theta}} d\lambda \\ &= \Gamma(\alpha+x) \left( \frac{\theta}{n\theta+1} \right)^{\alpha+x} \int_0^\infty \frac{1}{\Gamma(\alpha+x)} \left( \frac{\theta}{n\theta+1} \right)^{\alpha+x} \lambda^{\alpha+x-1} e^{-\frac{\lambda(n\theta+1)}{\theta}} d\lambda \\ &= \Gamma(\alpha+x) \left( \frac{\theta}{n\theta+1} \right)^{\alpha+x}, \end{aligned}$$

and

$$f_{\Lambda|{\mathbf{X}}={\mathbf{x}}}(\lambda) = \frac{1}{\Gamma(\alpha+x)} \left( \frac{\theta}{n\theta+1} \right)^{\alpha+x} \lambda^{\alpha+x-1} e^{-\frac{\lambda(n\theta+1)}{\theta}}.$$

**Example 9.3.3. Actuarial Exam Question.** You are given:

- The number of claims incurred in a month by any insured has a Poisson distribution with mean  $\lambda$ .
- The claim frequencies of different insured are iid.

- The prior distribution is gamma with pdf

$$f_{\Lambda}(\lambda) = \frac{(100\lambda)^6}{120\lambda} e^{-100\lambda}, \quad \lambda \in \mathbb{R}_+.$$

- The number of claims every month is distributed as follows:

Month	Number of Insured	Number of Claims
1	100	6
2	150	8
3	200	11
4	300	?

Calculate the expected number of claims in Month 4.

**Example Solution.** The likelihood function based on this policyholder's number of claims in Months 1, 2, and 3 is given by:

$$p_{\mathbf{X}|\Lambda=\lambda}(\mathbf{x}) = p_{X_1|\Lambda=\lambda}(6) p_{X_2|\Lambda=\lambda}(8) p_{X_3|\Lambda=\lambda}(11) \propto \lambda^{6+8+11} e^{-\lambda(100+150+200)}.$$

Because the prior distribution is gamma distributed with  $\alpha = 6$  and  $\theta = \frac{1}{100}$ , we know that the posterior distribution of parameter  $\lambda$  is also gamma distributed with shape parameter

$$\alpha + x = 6 + 6 + 8 + 11 = 31$$

and scale parameter

$$\frac{\theta}{n\theta + 1} = \frac{\frac{1}{100}}{(100 + 150 + 200)\frac{1}{100} + 1} = \frac{1}{550}.$$

The expected number of claim in Month 4 conditional on the information of Months 1, 2, and 3 is

$$E[E[X_4 | \Lambda = \lambda] | X_1, X_2, X_3] = E[300\lambda | X_1, X_2, X_3] = 300 E[\lambda | X_1, X_2, X_3],$$

and  $E[\lambda | X_1, X_2, X_3]$  is the expected value of the posterior distribution, which is given by

$$E[\lambda | X_1, X_2, X_3] = \frac{31}{550}.$$

Ultimately, this leads to an expected number of claim in Month 4 of  $300 \left( \frac{31}{550} \right) = \frac{930}{55} \approx 16.91$ .

### 9.3.3 The Normal–Normal Conjugate Family

The last conjugate family is the normal–normal family. Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  be a sample of iid normal random variables such that

$$f_{X_i|M=\mu}(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}\right), \quad x_i \in \mathbb{R}.$$

Further, to keep our focus on  $\mu$ , we will assume throughout our analysis that the variance parameter  $\sigma^2$  is known.<sup>13</sup> The likelihood function associated with this sample would therefore be given by

$$\begin{aligned} f_{\mathbf{X}|M=\mu}(\mathbf{x}) &= \prod_{i=1}^n f_{X_i|M=\mu}(x_i) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2}\right). \end{aligned}$$

A very natural prior distribution that matches the likelihood structure is unsurprisingly the normal distribution. Let us assume that the prior distribution for  $\mu$  is given by

$$f_M(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2}\frac{(\mu - \theta)^2}{\tau^2}\right),$$

where  $\theta$  is the mean parameter and  $\tau^2$  is the variance parameter. We can then easily show that the posterior distribution of  $\mu$  is also given by a normal distribution.

**Proposition 9.3.3. Normal–Normal Conjugate Family.** Consider a sample of  $n$  iid normals  $(X_1, X_2, \dots, X_n)$ , each with mean parameter  $\mu$  and variance parameter  $\sigma^2$  that is known. Further assume that the random variable associated with the mean,  $M$ , has a prior that is normally distributed with mean hyperparameter  $\theta$  and variance hyperparameter  $\tau^2$ . The posterior distribution of  $M$  is therefore given by

$$f_{M|\mathbf{X}=\mathbf{x}}(\mu) = \frac{1}{\sqrt{2\pi\left(\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}\right)}} \exp\left(-\frac{1}{2}\frac{\left(\mu - \left(\frac{x}{n}\frac{\tau^2}{n\tau^2+\sigma^2} + \theta\frac{\sigma^2}{n\tau^2+\sigma^2}\right)\right)^2}{\frac{\tau^2\sigma^2}{n\tau^2+\sigma^2}}\right),$$

---

<sup>13</sup>Conjugate families for the normal distribution with unknown  $\sigma^2$  can also be derived. For the sake of simplicity, we will only focus on the case with known variance parameter in this book.

where  $x = \sum_{i=1}^n x_i$ , which is a normal distribution with mean parameter

$$\frac{x}{n} \frac{n\tau^2}{n\tau^2 + \sigma^2} + \theta \frac{\sigma^2}{n\tau^2 + \sigma^2}$$

and variance parameter

$$\frac{\tau^2 \sigma^2}{n\tau^2 + \sigma^2}.$$

**Proof.** From Section 9.2.1, we know that

$$\begin{aligned} f_{M|\mathbf{X}=\mathbf{x}}(\mu) &= \frac{f_{\mathbf{X}|M=\mu}(\mathbf{x}) f_M(\mu)}{f_{\mathbf{X}}(\mathbf{x})} \\ &\propto \exp\left(-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma^2}\right) \exp\left(-\frac{1}{2} \frac{(\mu - \theta)^2}{\tau^2}\right) \\ &\propto \exp\left(-\frac{1}{2} \frac{\sum_{i=1}^n x_i^2 - 2\mu x + n\mu^2}{\sigma^2} - \frac{1}{2} \frac{\mu^2 - 2\mu\theta + \theta^2}{\tau^2}\right) \\ &\propto \exp\left(-\frac{1}{2} \frac{n\mu^2 - 2\mu x}{\sigma^2} - \frac{1}{2} \frac{\mu^2 - 2\mu\theta}{\tau^2}\right) \\ &\propto \exp\left(-\frac{1}{2} \frac{\mu^2(n\tau^2 + \sigma^2) - 2\mu\tau^2 x - 2\mu\sigma^2\theta}{\tau^2\sigma^2}\right) \\ &\propto \exp\left(-\frac{1}{2} \frac{\mu^2 - 2\mu \left(x \frac{\tau^2}{n\tau^2 + \sigma^2} + \theta \frac{\sigma^2}{n\tau^2 + \sigma^2}\right)}{\frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}}\right) \\ &\propto \frac{1}{\sqrt{2\pi \left(\frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}\right)}} \exp\left(-\frac{1}{2} \frac{\left(\mu - \left(x \frac{n\tau^2}{n\tau^2 + \sigma^2} + \theta \frac{\sigma^2}{n\tau^2 + \sigma^2}\right)\right)^2}{\frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}}\right), \end{aligned}$$

where  $x = \sum_{i=1}^n x_i$ .

The prior distribution hyperparameters and posterior distribution parameters can be interpreted in the normal–normal conjugate family:

- For the prior,  $\theta$  represents the *a priori* value of the mean parameter, and  $\tau^2$  is related to the precision of that prior mean (i.e., the larger the value, the less precise the prior mean is, and vice versa).
- For the posterior, the new mean parameter is a weighted average between the prior mean parameter  $\theta$  and the sample mean  $\frac{x}{n}$ . The new variance parameter is informed by the prior variability  $\tau^2$  and the variability of the data  $\sigma^2$ .

**Example 9.3.4. Impact of Normal Prior on Posterior.** Assume the following observed automobile claims for a small portfolio of policies:

$$1050, \quad 1250, \quad 1550, \quad 2600, \quad 5350, \quad 10200.$$

Further assume that the logarithm of the claim amount follows a normal distribution with parameters  $\mu$  and  $\sigma^2 = 1$ . Find the posterior distribution of the mean parameter  $\mu$  for a normal prior distribution where  $\theta = 7$ . Consider different values of  $\tau^2$ ; that is,  $\tau^2 = 0.1$ ,  $\tau^2 = 1$ , and  $\tau^2 = 10$ . Figure 9.10 shows the pdf of these three prior distributions.

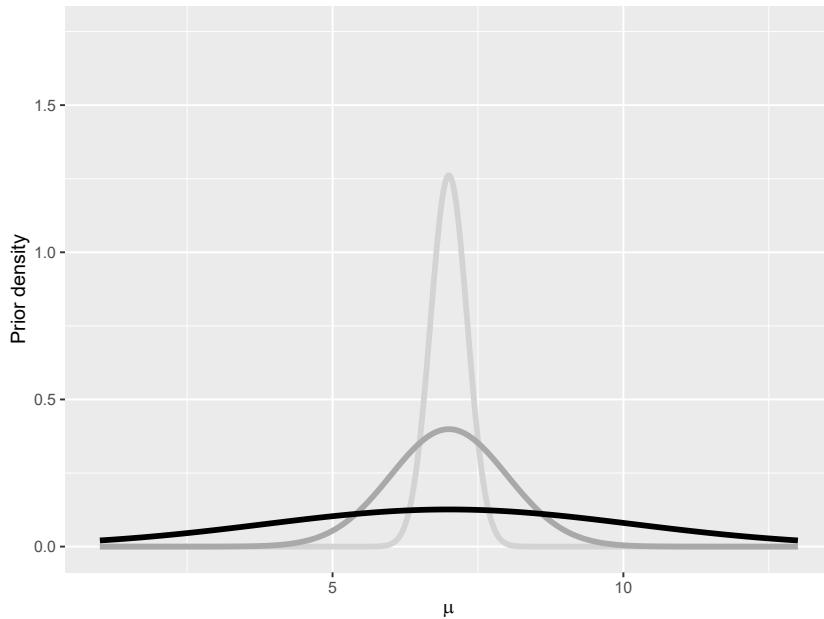


FIGURE 9.10: Normal prior densities:  $\tau^2 = 0.1$  (light gray),  $\tau^2 = 1$  (gray), and  $\tau^2 = 10$  (black)

**Example Solution.** Using the results of Proposition 9.3.3, we can obtain the following posterior distributions:

```

xi <- c(1050, 1250, 1550, 2600, 5350, 10200)
x <- sum(log(xi))
n <- length(xi)
sigma2 <- 1

mean1 <- theta * (sigma2/(n * tau21 + sigma2)) + x/n * ((n * tau21)/(n * tau21 +
sigma2))

```

```

mean2 <- theta * (sigma2/(n * tau22 + sigma2)) + x/n * ((n * tau22)/(n * tau22 +
  sigma2))
mean3 <- theta * (sigma2/(n * tau23 + sigma2)) + x/n * ((n * tau23)/(n * tau23 +
  sigma2))

var1 <- (tau21 * sigma2)/(n * tau21 + sigma2)
var2 <- (tau22 * sigma2)/(n * tau22 + sigma2)
var3 <- (tau23 * sigma2)/(n * tau23 + sigma2)

posterior1 <- dnorm(xs, mean = mean1, sd = sqrt(var1))
posterior2 <- dnorm(xs, mean = mean2, sd = sqrt(var2))
posterior3 <- dnorm(xs, mean = mean3, sd = sqrt(var3))

dataposterior <- data.frame(x = xs, y1 = posterior1, y2 = posterior2, y3 = posterior3)

ggplot(dataposterior, aes(x = x, y = y1)) + geom_line(color = "lightgray", lwd = 1.5) +
  geom_line(aes(y = y2), color = "darkgray", lwd = 1.5) + geom_line(aes(y = y3),
  color = "black", lwd = 1.5) + xlim(1, 13) + ylim(0, 1.75) + xlab(expression(italic(mu))) +
  ylab("Posterior density")

```

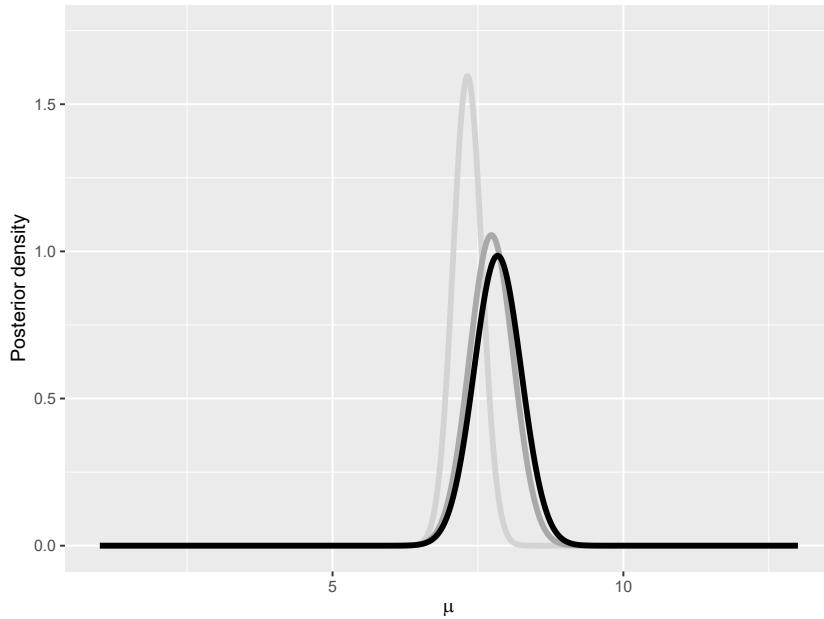


FIGURE 9.11: Posterior densities based on three different priors:  $\tau^2 = 0.1$  (light gray),  $\tau^2 = 1$  (gray), and  $\tau^2 = 10$  (black)

---

Interestingly, as shown in Example 9.3.4, the prior distribution can have some impact on the final posterior distribution. When the prior assumption about the mean is very precise, having a few data points do not create a huge gap between the prior and the posterior (see the light gray curves in Figures 9.10

and 9.11). When the prior is very imprecise, on the other hand, then the data are allowed to speak, and the posterior can be quite different from the prior distribution.

#### 9.3.4 Criticism of Conjugate Family Models

While conjugate family models have some advantages, such as ease of interpretation and computational simplicity, they also have some limitations:

1. Conjugate families are oftentimes chosen for their mathematical convenience rather than their ability to accurately model the data under study. This can lead to models that are too simplistic and lack the flexibility needed to model real-world phenomena.
2. Conjugate family models rely on the choice of prior distribution, and different choices of possibly non-conjugate priors can lead to very different posterior distributions.
3. Conjugate family models are only applicable to a narrow range of problems, which limit their usefulness in practical applications.

It is important to note that while conjugate family models have their limitations, they can still be useful in certain situations, especially when the assumptions of the model are well understood and the data are relatively simple.

---

## 9.4 Posterior Simulation

---

In Section 9.4, you learn how to:

- Use the standard computational tools for Bayesian statistics.
  - Diagnose Markov chain convergence.
- 

### 9.4.1 Introduction to Markov Chain Monte Carlo Methods

Sometimes, using conjugate family models is ill-suited for the problem at hand, and more complicated priors need to be selected. Under other circumstances, complex models involve many parameters making the posterior distribution intractable. In these cases, the posterior distribution of the parameters will not

have a closed-form solution, generally speaking, and will need to be estimated via numerical methods.

A common way to generate draws of the parameter posterior distribution is to create Markov chains for which their stationary distributions—the probability distribution that remains unchanged when the Markov chain has reached a state where the transition probabilities no longer evolve over time—correspond to the posterior of interest. These Markov chain-based methods are known as Markov chain Monte Carlo (MCMC) methods in the literature. This section provides a brief overview of these methods and of their uses. We do not intend to give much of the theory behind these methods, which would require a deep understanding of Markov chains and their theory.<sup>14</sup> Instead, we focus on their applications in insurance and loss modeling. Specifically, in the next two subsections, we introduce the two most common MCMC methods; that is, the Gibbs sampler of [Gelfand and Smith \(1990\)](#) and the Metropolis–Hastings algorithm of [Hastings \(1970\)](#) and [Metropolis et al. \(1953\)](#).

#### 9.4.2 The Gibbs Sampler

As mentioned above, sometimes, we cannot use conjugate families. In other cases where the parameter space is large, it can be very hard to find the marginal likelihood  $f_{\mathbf{X}}(\mathbf{x})$  (also known as the normalizing constant); that is, assuming that the model parameters are given by  $\boldsymbol{\theta} = [\theta_1 \dots \theta_2 \dots \theta_k]$  and contains  $k$  parameters, the marginal likelihood given by

$$f_{\mathbf{X}}(\mathbf{x}) = \int \int \dots \int f_{\mathbf{X}|\boldsymbol{\Theta}=\boldsymbol{\theta}}(\mathbf{x}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) d\theta_1 d\theta_2 \dots d\theta_k$$

is hard to compute even when using typical quadrature-based rules, especially if  $k$  is large.

Fortunately, under very mild regularity conditions, samples of the joint estimates of parameters can be obtained by sequentially sampling each parameter individually and by keeping all the other parameters constant. To do so, the distribution of any given parameter conditional on all the other parameters (and the data) needs to be known. These distributions are known as full conditional distributions; that is,

$$f_{\Theta_i | \mathbf{X}=\mathbf{x}, \boldsymbol{\Theta}_{\setminus i}=\boldsymbol{\theta}_{\setminus i}}(\theta_i),$$

for parameter  $\theta_i$ , where  $\boldsymbol{\theta}_{\setminus i}$  represents all parameters except for the  $i^{\text{th}}$  one, and  $\boldsymbol{\Theta}_{\setminus i}$  is the random variable associated with this set of parameters.

---

<sup>14</sup>For an overview of the theory behind MCMC methods, see [Robert and Casella \(1999\)](#).

The full conditional distribution is an important building block in Gibbs sampling. Indeed, if one can obtain each parameter's distribution conditional on having the value of all the other parameters in closed form, then it is possible to generate samples for each parameter. Specifically, starting from an arbitrary set of starting values  $\boldsymbol{\theta}^{(0)} = [\theta_1^{(0)} \quad \theta_2^{(0)} \quad \dots \quad \theta_k^{(0)}]$ , samples for each parameter can be generated by performing the following steps for  $m = 1, 2, \dots, M$ :

1. Draw  $\theta_1^{(m)}$  from  $f_{\Theta_1 | \mathbf{X}=\mathbf{x}, \Theta_2=\theta_2^{(m-1)}, \dots, \Theta_k=\theta_k^{(m-1)}}(\theta_1)$ .
  2. Draw  $\theta_2^{(m)}$  from  $f_{\Theta_2 | \mathbf{X}=\mathbf{x}, \Theta_1=\theta_1^{(m)}, \Theta_3=\theta_3^{(m-1)}, \dots, \Theta_k=\theta_k^{(m-1)}}(\theta_2)$ .
  3. Draw  $\theta_3^{(m)}$  from  $f_{\Theta_3 | \mathbf{X}=\mathbf{x}, \Theta_1=\theta_1^{(m)}, \Theta_2=\theta_2^{(m)}, \Theta_4=\theta_4^{(m-1)}, \dots, \Theta_k=\theta_k^{(m-1)}}(\theta_3)$ .
- $\vdots$
- $k.$  Draw  $\theta_k^{(m)}$  from  $f_{\Theta_k | \mathbf{X}=\mathbf{x}, \Theta_1=\theta_1^{(m)}, \dots, \Theta_{k-1}=\theta_{k-1}^{(m)}}(\theta_k)$ .

The sample, especially at first, will depend on the initial values,  $\boldsymbol{\theta}^{(0)}$ , and it might take some time until the sampler can get to the stationary distribution. For this reason, in practice, experimenters discard the first  $M^*$  iterations to make sure their analysis is not impacted by the choice of initial parameter; this initial period of discarded sample is known as the burn-in period.

The rest of the sample—the remaining  $M - M^*$  iterations—is kept to estimate the posterior distribution and any quantities of interest.

#### Application to Bayesian Linear Regression

In statistics and in its most simple form, a linear regression is an approach for modeling the relationship between a scalar response and an explanatory variable. The former quantity is denoted by  $x_i$  for  $i \in \{1, \dots, n\}$ , and the latter quantity is denoted by  $z_i$  for  $i \in \{1, \dots, n\}$  in this chapter. Mathematically, we can write this relationship as

$$x_i = \alpha + \beta z_i + \varepsilon_i,$$

where  $\varepsilon_i$  is a disturbance term that captures the potential for errors in the linear relationship. This error term is typically assumed to be normally distributed with mean zero and variance  $\sigma^2$ .

In general, the coefficients  $\alpha$  and  $\beta$  are unknown and need to be estimated. The experimenter can rely on Bayesian statistics to find out the posterior distribution of the parameters  $\alpha$  and  $\beta$  along with that of  $\sigma^2$ . For the rest of the subsection, we investigate a specific application of Gibbs sampling to the context of linear regression.

We begin by computing the likelihood function conditional on the parameter values:

$$\begin{aligned} f_{\mathbf{X} | A=\alpha, B=\beta, \Sigma^2=\sigma^2}(\mathbf{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \alpha - \beta z_i)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (x_i - \alpha - \beta z_i)^2}{2\sigma^2}\right), \end{aligned}$$

which is the first building block to construct our posterior distribution.

Then, we need a prior distribution, which could be informative, weakly informative, or noninformative. In this application, we select a prior that allows us to obtain each parameter's full conditional distribution in closed form. Specifically, we use a normal distribution for  $\alpha$  and  $\beta$ , and an inverse gamma distribution for  $\sigma^2$  with shape parameter  $\frac{n_\sigma}{2}$  and scale parameter  $\frac{\theta_\sigma}{2}$ , where

$$\begin{aligned} f_A(\alpha) &= \frac{1}{\sqrt{2\pi\tau_\alpha^2}} \exp\left(-\frac{1}{2} \frac{(\alpha - \theta_\alpha)^2}{\tau_\alpha^2}\right), \\ f_B(\beta) &= \frac{1}{\sqrt{2\pi\tau_\beta^2}} \exp\left(-\frac{1}{2} \frac{(\beta - \theta_\beta)^2}{\tau_\beta^2}\right), \\ f_{\Sigma^2}(\sigma^2) &= \frac{(\theta_\sigma/2)^{n_\sigma/2}}{\Gamma(n_\sigma/2)} \left(\frac{1}{\sigma^2}\right)^{n_\sigma/2+1} \exp\left(-\frac{\theta_\sigma/2}{\sigma^2}\right). \end{aligned}$$


---

**Proposition 9.4.1. Full Conditional Distributions of Bayesian Linear Regression Parameters.** Consider a sample of  $n$  observations  $\mathbf{x} = (x_1, \dots, x_n)$  for which

$$x_i = \alpha + \beta z_i + \varepsilon_i,$$

where  $\varepsilon_i$  is normally distributed with mean zero and variance  $\sigma^2$ . The full conditional distributions of parameters  $\alpha$ ,  $\beta$ , and  $\sigma^2$  are given by the following expressions:

$$\begin{aligned} A &\sim \text{Normal}\left(\frac{1}{n} \left(\sum_{i=1}^n x_i - \beta z_i\right) \frac{n\tau_\alpha^2}{n\tau_\alpha^2 + \sigma^2} + \theta_\alpha \frac{\sigma^2}{n\tau_\alpha^2 + \sigma^2}, \frac{\tau_\alpha^2 \sigma^2}{n\tau_\alpha^2 + \sigma^2}\right), \\ B &\sim \text{Normal}\left(\frac{1}{n} \left(\sum_{i=1}^n z_i (x_i - \alpha)\right) \frac{n\tau_\beta^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2} + \theta_\beta \frac{\sigma^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2}, \frac{\tau_\beta^2 \sigma^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2}\right), \\ \Sigma^2 &\sim \text{Inverse Gamma}\left(\frac{n_\sigma + n}{2}, \frac{\theta_\sigma + \sum_{i=1}^n (y_i - \alpha - \beta z_i)^2}{2}\right), \end{aligned}$$

respectively, assuming the prior distributions mentioned above.

**Proof.** From Section refS:Sec92, we know that

$$f_{A,B,\Sigma^2|\mathbf{X}=\mathbf{x}}(\alpha, \beta, \sigma^2) \propto f_{\mathbf{X}|A=\alpha, B=\beta, \Sigma^2=\sigma^2}(\mathbf{x}) f_A(\alpha) f_B(\beta) f_{\Sigma^2}(\sigma^2),$$

which is useful to derive the full conditional distributions of  $\alpha$ ,  $\beta$ , and  $\sigma^2$ .

Let us begin with  $\alpha$ :

$$\begin{aligned} & f_{A|\mathbf{X}=\mathbf{x}, B=\beta, \Sigma^2=\sigma^2}(\alpha) \\ & \propto \exp\left(-\frac{1}{2} \frac{\sum_{i=1}^n (x_i - \alpha - \beta z_i)^2}{\sigma^2}\right) \exp\left(-\frac{1}{2} \frac{(\alpha - \theta_\alpha)^2}{\tau_\alpha^2}\right) \\ & \propto \exp\left(-\frac{1}{2} \left( \frac{n\alpha^2 - 2\alpha \sum_{i=1}^n (x_i - \beta z_i)}{\sigma^2} + \frac{\alpha^2 - 2\alpha \theta_\alpha}{\tau_\alpha^2} \right)\right) \\ & \propto \exp\left(-\frac{1}{2} \left( \frac{\alpha^2 - 2\alpha \left( \frac{1}{n} (\sum_{i=1}^n x_i - \beta z_i) \frac{n\tau_\alpha^2}{n\tau_\alpha^2 + \sigma^2} + \theta_\alpha \frac{\sigma^2}{n\tau_\alpha^2 + \sigma^2} \right)}{\frac{\tau_\alpha^2 \sigma^2}{n\tau_\alpha^2 + \sigma^2}} \right)\right) \\ & \propto \frac{1}{\sqrt{2\pi \left( \frac{\tau_\alpha^2 \sigma^2}{n\tau_\alpha^2 + \sigma^2} \right)}} \exp\left(-\frac{1}{2} \left( \frac{\left( \alpha - \left( \frac{1}{n} (\sum_{i=1}^n x_i - \beta z_i) \frac{n\tau_\alpha^2}{n\tau_\alpha^2 + \sigma^2} + \theta_\alpha \frac{\sigma^2}{n\tau_\alpha^2 + \sigma^2} \right) \right)^2}{\frac{\tau_\alpha^2 \sigma^2}{n\tau_\alpha^2 + \sigma^2}} \right)\right) \end{aligned}$$

which is a normal distribution with mean parameter

$$\frac{1}{n} \left( \sum_{i=1}^n x_i - \beta z_i \right) \frac{n\tau_\alpha^2}{n\tau_\alpha^2 + \sigma^2} + \theta_\alpha \frac{\sigma^2}{n\tau_\alpha^2 + \sigma^2}$$

and variance parameter

$$\frac{\tau_\alpha^2 \sigma^2}{n\tau_\alpha^2 + \sigma^2}.$$

The derivation to obtain the full conditional distribution of  $\beta$  is similar to that of  $\alpha$ :

$$\begin{aligned}
& f_{B|\mathbf{X}=\mathbf{x}, A=\alpha, \Sigma^2=\sigma^2}(\beta) \\
& \propto \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^n (x_i - \alpha - \beta z_i)^2}{\sigma^2}\right) \exp\left(-\frac{1}{2}\frac{(\beta - \theta_\beta)^2}{\tau_\beta^2}\right) \\
& \propto \exp\left(-\frac{1}{2}\left(\frac{\beta^2 \sum_{i=1}^n z_i^2 - 2\beta \sum_{i=1}^n z_i(x_i - \alpha)}{\sigma^2} + \frac{\beta^2 - 2\beta\theta_\beta}{\tau_\beta^2}\right)\right) \\
& \propto \exp\left(-\frac{1}{2}\left(\frac{\beta^2 - 2\beta\left(\frac{1}{n}(\sum_{i=1}^n z_i(x_i - \alpha))\right)\frac{n\tau_\beta^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2} + \theta_\beta \frac{\sigma^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2}}{\frac{\sigma_\beta^2 \sigma^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2}}\right)\right) \\
& \propto \frac{1}{\sqrt{2\pi\left(\frac{\tau_\beta^2 \sigma^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2}\right)}} \\
& \times \exp\left(-\frac{1}{2}\left(\frac{\left(\beta - \left(\frac{1}{n}(\sum_{i=1}^n z_i(x_i - \alpha))\right)\frac{n\tau_\beta^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2} + \theta_\beta \frac{\sigma^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2}\right)\right)^2}{\frac{\tau_\beta^2 \sigma^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2}}\right)
\end{aligned}$$

which is a normal distribution with mean parameter

$$\frac{1}{n} \left( \sum_{i=1}^n z_i (x_i - \alpha) \right) \frac{n\tau_\beta^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2} + \theta_\beta \frac{\sigma^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2}$$

and variance parameter

$$\frac{\tau_\beta^2 \sigma^2}{\tau_\beta^2 \sum_{i=1}^n z_i^2 + \sigma^2}.$$

Finally, we apply the same logic to the variance parameter,  $\sigma^2$ :

$$\begin{aligned}
 & f_{\Sigma^2 | \mathbf{X} = \mathbf{x}, A = \alpha, B = \beta}(\sigma^2) \\
 & \propto (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^n (x_i - \alpha - \beta z_i)^2}{\sigma^2}\right) \left(\frac{1}{\sigma^2}\right)^{n_\sigma/2+1} \exp\left(-\frac{\theta_\sigma/2}{\sigma^2}\right) \\
 & \propto \exp\left(-\frac{1}{2}\frac{\theta_\sigma + \sum_{i=1}^n (x_i - \alpha - \beta z_i)^2}{\sigma^2}\right) \left(\frac{1}{\sigma^2}\right)^{(n_\sigma+n)/2+1} \\
 & \propto \frac{\left(\frac{\theta_\sigma + \sum_{i=1}^n (x_i - \alpha - \beta z_i)^2}{2}\right)^{(n_\sigma+n)/2}}{\Gamma((n_\sigma + n)/2)} \left(\frac{1}{\sigma^2}\right)^{(n_\sigma+n)/2+1} \exp\left(-\frac{\frac{\theta_\sigma + \sum_{i=1}^n (x_i - \alpha - \beta z_i)^2}{2}}{\sigma^2}\right),
 \end{aligned}$$

which is an inverse gamma distribution with shape parameter  $\frac{n_\sigma+n}{2}$  and scale parameter

$$\frac{\theta_\sigma + \sum_{i=1}^n (x_i - \alpha - \beta z_i)^2}{2}.$$

---

We now apply the Gibbs sampler on *real* data. The example will use motorcycle insurance data from Wasa, a Swedish insurance company, taken from `data0hlsson` of the R package `insuranceData`; see [Wolny-Dominiak and Trzesiok \(2014\)](#) for more details.

```
library("insuranceData")
data(data0hlsson)
```

This dataset contains information about the number of motorcycle accidents, their claim cost, and some risk factors (e.g., the age of the driver, the age of the vehicle, the geographic zone).

---

**Example 9.4.1. Bayesian Linear Regression.** You wish to understand the relationship between the age of the driver and the (logarithm of the) claim cost. Let  $x_i$  be the logarithm of the  $i^{\text{th}}$  claim cost and  $z_i$  be the age associated with the  $i^{\text{th}}$  claim. Further assume the following linear relationship between the two quantities:

$$x_i = \alpha + \beta z_i + \varepsilon_i,$$

where  $\varepsilon_i$  is normally distributed with mean zero and variance  $\sigma^2$ . Find the posterior density of the three parameters  $\alpha$ ,  $\beta$ , and  $\sigma^2$  using the Gibbs sampler.

**Example Solution.** Let us begin by visualizing the data. Figure 9.12 reports the logarithm of the claim cost as a function of the driver's age. At first sight, it seems that the relationship between the claim cost and age is negative, so we should expect a negative  $\beta$ , generally speaking.

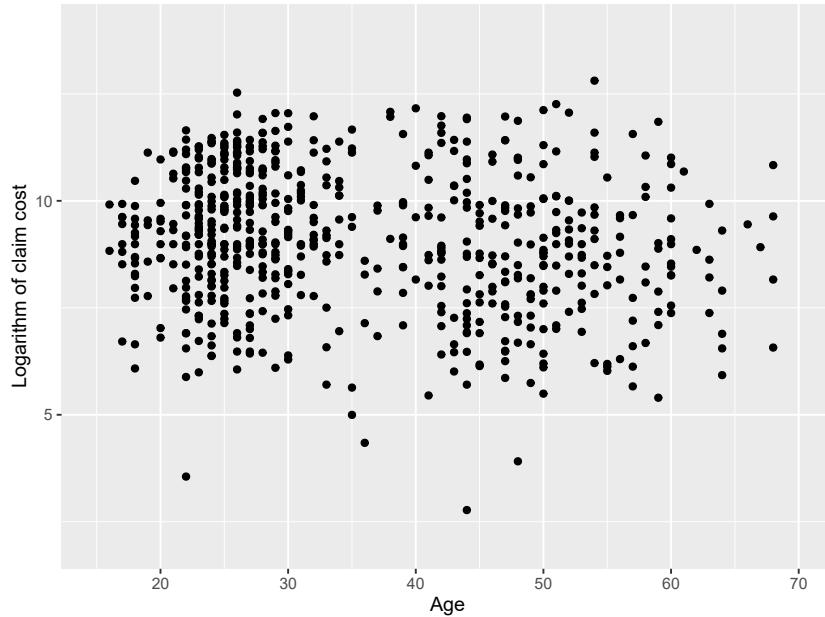


FIGURE 9.12: Logarithm of the claim cost as a function of the driver's age

Let us now turn to Bayesian computation via Gibbs sampling to find the posterior distribution of the three parameters of interest. We will use 10,000 iterations and discard the first 5,000 iterations (i.e., burn-in period). For our prior distributions, we use weakly informative priors by setting  $\theta_\alpha = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$ ,  $\theta_\beta = 0$ ,  $\tau_\alpha^2 = \tau_\beta^2 = 10$ ,  $n_\sigma = 1$ , and  $\theta_\sigma = 0.1$ . The initial values of the parameters are set to:  $\alpha^{(0)} = \bar{x}$ ,  $\beta^{(0)} = 0$ , and  $\sigma^{2(0)} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ .

```
set.seed(1)
library("nimble")
dataOhlsson <- dataOhlsson[dataOhlsson$skadkost > 0, ]
dataOhlsson$logskadkost <- log(dataOhlsson$skadkost)

x <- dataOhlsson$logskadkost
z <- dataOhlsson$agarald
```

```

n <- length(x)
M <- 10000
Mstar <- 5000
thetaa <- mean(x)
tau2a <- 10
thetab <- 0
tau2b <- 10
nsigma <- 1
thetasigma <- 0.1

alphas <- rep(NA, M + 1)
betas <- rep(NA, M + 1)
sigma2s <- rep(NA, M + 1)

alphas[1] <- mean(x)
betas[1] <- 0
sigma2s[1] <- var(x)

for (m in 2:(M + 1)) {
  # Generate alpha
  den_alpha <- n * tau2a + sigma2s[m - 1]
  mean_alpha <- (1/n) * (sum(x - betas[m - 1] * z)) * (n * tau2a)/den_alpha + thetaa *
    sigma2s[m - 1]/den_alpha
  var_alpha <- tau2a * sigma2s[m - 1]/den_alpha

  alphas[m] <- rnorm(1, mean = mean_alpha, sd = sqrt(var_alpha))

  # Generate beta
  den_beta <- tau2b * sum(z^2) + sigma2s[m - 1]
  mean_beta <- (1/n) * (sum(z * (x - alphas[m]))) * (n * tau2b)/den_beta + thetab *
    sigma2s[m - 1]/den_beta
  var_beta <- tau2b * sigma2s[m - 1]/den_beta

  betas[m] <- rnorm(1, mean = mean_beta, sd = sqrt(var_beta))

  # Generate sigma^2
  shape_sigma <- (nsigma + n)/2
  scale_sigma <- (thetasigma + sum((x - alphas[m] - betas[m] * z)^2))/2

  sigma2s[m] <- rinvgamma(1, shape = shape_sigma, scale = scale_sigma)
}

```

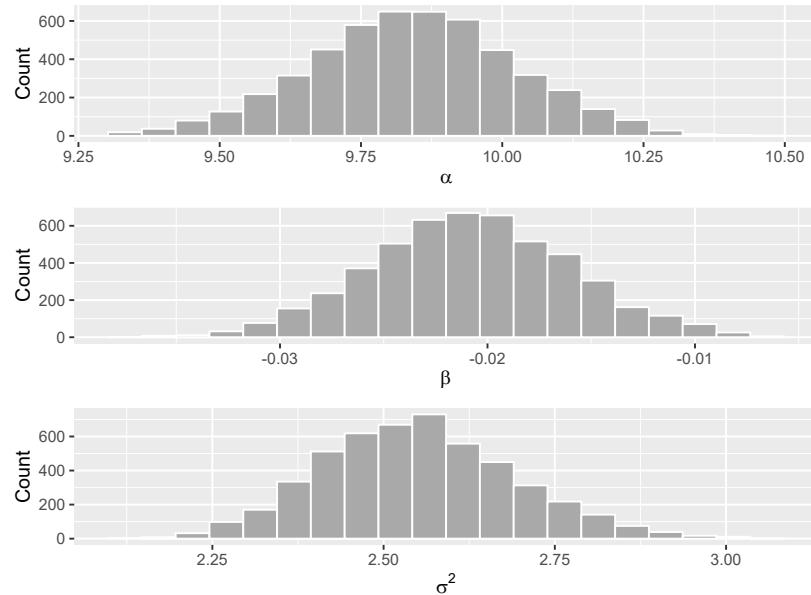
Once we have the posterior parameter samples, we can get multiple quantities of interest. For instance, the posterior mean of parameters  $\alpha$ ,  $\beta$ , and  $\sigma^2$  are 9.843,  $-0.0208$ , and 2.551, respectively. These posterior means are obtained by simply taking the sample means of the respective posterior draws; that is, these are Monte Carlo estimates of the posterior means.

The posterior mean for coefficient alpha is 9.84259435

The posterior mean for coefficient beta is -0.0207847772

The posterior mean for the variance parameter is 2.55090494

We can also get histograms of the posterior distribution for  $\alpha$ ,  $\beta$ , and  $\sigma^2$ ; Figure 9.13 reports histograms for the three parameters. The uncertainty around each parameter is very small.



**FIGURE 9.13: Histogram of the posterior distribution for parameters  $\alpha$  (top panel),  $\beta$  (middle panel), and  $\sigma^2$  (bottom panel)**

The top panel of Figure 9.14 reports a plot of the post-burn-in values of  $\alpha$  as a function of the iteration number; this type of plot is known as a trace plot in the literature. These samples are not impacted by the initial parameter value that was selected. Indeed, after about 20–30 iterations, the posterior parameter values obtained by the Gibbs sampler are very close to their posterior means. For instance, the bottom panel of Figure reffig:Fig914 shows a plot of the first 50 values of  $\alpha$  as a function of the iteration number.

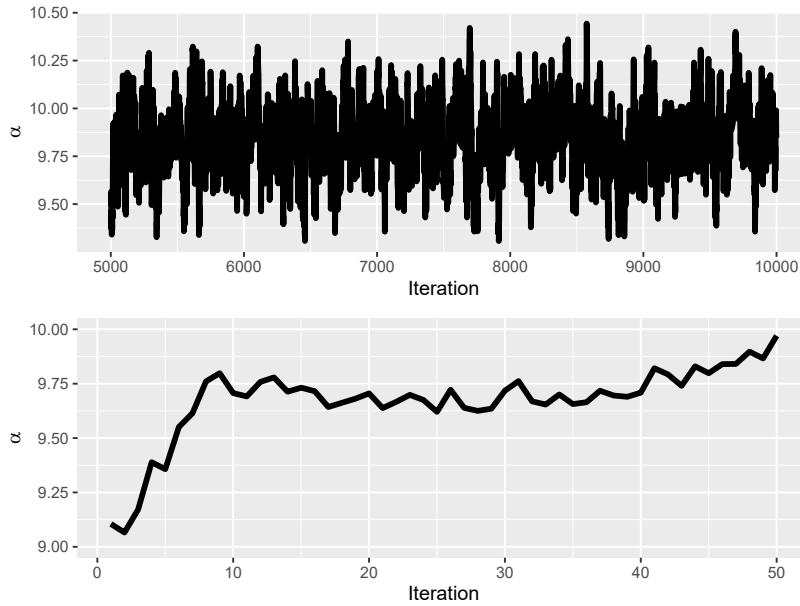


FIGURE 9.14: Trace plot of  $\alpha$  for the post-burn-in iterations (top panel) and for the first 50 iterations (bottom panel)

#### 9.4.3 The Metropolis–Hastings Algorithm

Gibbs sampling works well when the full conditional distribution for each parameter in the model can be found and is of a common form. This, unfortunately, is not always possible, meaning that we need to rely on other computational tools to find the posterior distribution of the parameters. One very popular method that copes with the shortcomings of Gibbs' method is the Metropolis–Hastings sampler.

Let us assume that the current value of the first model parameter is  $\theta_1^{(0)}$ . From this current value, we now wish to find a new value for this parameter. To do so, we propose a new value for this parameter,  $\theta_1^*$ , from a candidate (or proposal) density  $q(\theta_1^* | \theta_1^{(0)})$ . Since this proposal has nothing to do with the posterior distribution of the parameter, we should not keep all candidates in our final sample—we only accept those samples that are representative of the posterior distribution of interest. To determine whether we accept or reject the candidate, we compute a so-called acceptance ratio  $\alpha(\theta_1^{(0)}, \theta_1^*)$  using

$$\alpha(\theta_1^{(0)}, \theta_1^*) = \frac{h(\theta_1^*) q(\theta_1^{(0)} | \theta_1^*)}{h(\theta_1^{(1)}) q(\theta_1^* | \theta_1^{(0)})}$$

where

$$h(\theta_1) = f_{\mathbf{x} | \Theta_1 = \theta_1, \Theta_{\setminus 1} = \boldsymbol{\theta}_{\setminus 1}}(\mathbf{x}) f_{\Theta_1, \Theta_{\setminus 1}}(\theta_1, \boldsymbol{\theta}_{\setminus 1})$$

and  $\boldsymbol{\theta}_{\setminus 1}$  represents all parameters except for the first one. Then, we accept the proposed value  $\theta_1^*$  with probability  $\alpha(\theta_1^{(0)}, \theta_1^*)$  and reject it with probability  $1 - \alpha(\theta_1^{(0)}, \theta_1^*)$ . Specifically,

$$\theta_1^{(1)} = \begin{cases} \theta_1^* & \text{with probability } \alpha(\theta_1^{(0)}, \theta_1^*) \\ \theta_1^{(0)} & \text{with probability } 1 - \alpha(\theta_1^{(0)}, \theta_1^*) \end{cases}$$

We can repeat the same process for all other parameters to obtain  $\theta_2^{(1)}$  to  $\theta_k^{(1)}$ , while replacing the parameters  $\boldsymbol{\theta}_{\setminus i}$  by their most current values in the chain. Once we have updated all values, we can repeat this process for all  $m$  in  $\{2, 3, \dots, M\}$ , similar to the iterative process used in the Gibbs sampler.<sup>15</sup>

**Special Case: Symmetric Proposal Distribution.** If a proposal distribution is symmetric, then

$$q(\theta_i^{(m)} | \theta_i^*) = q(\theta_i^* | \theta_i^{(m)}) ,$$

and those terms cancel out, leaving

$$\alpha(\theta_i^{(m)}, \theta_1^*) = \frac{h(\theta_i^*)}{h(\theta_i^{(m)})} .$$

This special case is called the *Metropolis algorithm*.

The Metropolis–Hastings sampler requires a lot of fine-tuning, generally speaking, because the experimenter needs to select a proposal distribution for each parameter. A common approach is to assume a normal proposal distribution centered at the previous value; that is,

$$\Theta_i^* \sim \text{Normal}(\theta_i^{(m-1)}, \delta_i^2) ,$$

at step  $m$ , where  $\delta_i^2$  is the variance of the  $i^{\text{th}}$  parameter's proposal distribution.

**Example 9.4.2. Impact of Proposal Density on the Acceptance Rate.** Assume that each policyholder's claim count (frequency) is distributed as a Poisson random variable such that

$$p_{N_i | \Lambda=\lambda}(n_i) = \frac{\lambda^{n_i} e^{-\lambda}}{n_i!} ,$$

---

<sup>15</sup>The Gibbs sampler can be seen as a special case of the more general Metropolis–Hastings algorithm. Specifically, with Gibbs' method, all proposals are automatically accepted; that is,  $\alpha(\theta_1^{(0)}, \theta_1^*) = 1$ .

where  $n_i$  is the number of claims associated with the  $i^{\text{th}}$  policyholder. Further assume a noninformative, flat prior over  $[0, \infty]$ ; that is,

$$f_{\Lambda}(\lambda) \propto 1, \quad \lambda \in [0, \infty].$$

Find the posterior distribution of the parameter using 1,000 iterations of the Metropolis–Hastings sampler assuming the claim count data of the Singapore Insurance Data (see Example 9.1.4 for more details). Use a normal proposal with small ( $1 \times 10^{-7}$ ), moderate ( $1 \times 10^{-4}$ ), and large ( $1 \times 10^{-1}$ ) values as the proposal variance  $\delta$  in your tests and comment on the differences.

**Example Solution.** Starting from the likelihood function and the prior distribution, we have that

$$h(\lambda) \propto \prod_{i=1}^N \frac{\lambda^{n_i} e^{-\lambda}}{n_i!}.$$

You can learn more about the R code for this example at the online version of this book, [Actuarial Community \(2025\)](#).

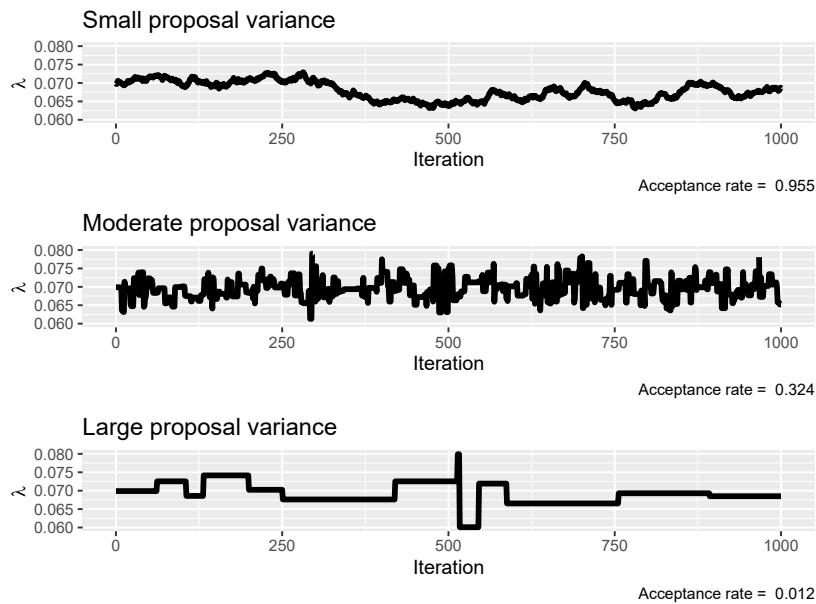


FIGURE 9.15: Trace plots based on three different proposals:  $\sigma^2 = 1 \times 10^{-7}$  (top panel),  $\sigma^2 = 1 \times 10^{-4}$  (middle panel), and  $\sigma^2 = 1 \times 10^{-1}$  (bottom panel)

Different variance parameters lead to different results. In this example, if  $\delta^2$  is too small, then the experimenter tends to draw samples that are very similar from one iteration to the other. This increases the acceptance rate (i.e., the rate at which we accept the proposal), but also means that the chain is travelling slowly around the posterior distribution. This ultimately imply that it will take longer chains to visit the whole posterior distribution. One way to see this issue in practice is by computing autocorrelation coefficients for the sample of parameter (more details on this in Section 9.4.4). The top panel of Figure 9.1.5 indeed shows this strong autocorrelation and slow travelling around the posterior distribution.

On the other hand, if  $\delta^2$  is too large, then the proposal are seldom accepted, and the chain will tend to stick—exhibiting long period for which the chain stays constant. For instance, the case with large proposal variance above leads to an acceptance rate of 1.2 percent, which is very low. The bottom panel of Figure 9.1.5 reports this issue.

The moderate proposal variance case reports an acceptance rate of 32.4 percent, which is not too high nor too low. The general behavior of this chain resembles that of a hairy caterpillar—a good sign—meaning that the mixing seems adequate and that we accept a decent amount of proposed values.

Finding the right proposal variance values for problems of interest requires some fine-tuning. As a general guideline, experimenters should target acceptance rates between 20 and 50 percent.

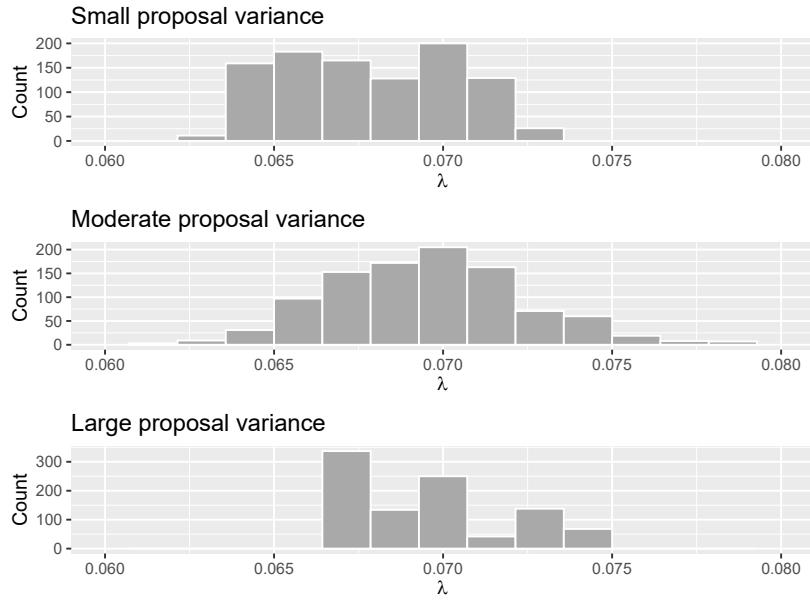
Using the wrong proposal distribution can have an impact on the computational efficiency of the Metropolis–Hastings algorithm, as shown in Figure 9.16. A small variance takes a long time to travel throughout the posterior distribution, whereas a large variance tends to stick.

---

**Example 9.4.3. Impact of Initial Parameters.** Consider the motorcycle insurance data from Wasa used in Example 9.4.1. We wish to model the claim amount from motorcycle losses with a gamma distribution; that is,

$$f_{X_i | \Theta=\theta, A=\alpha}(x_i) = \frac{1}{\theta^\alpha \Gamma(\alpha)} x_i^{\alpha-1} e^{-\frac{x_i}{\theta}},$$

where  $x_i$  is the  $i^{\text{th}}$  claim amount. We assume that the prior distributions for



**FIGURE 9.16:** Posterior densities based on three different proposals:  $\sigma^2 = 1 \times 10^{-7}$  (top panel),  $\sigma^2 = 1 \times 10^{-4}$  (middle panel), and  $\sigma^2 = 1 \times 10^{-1}$  (bottom panel)

both  $\theta$  and  $\alpha$  are noninformative and flat; that is,

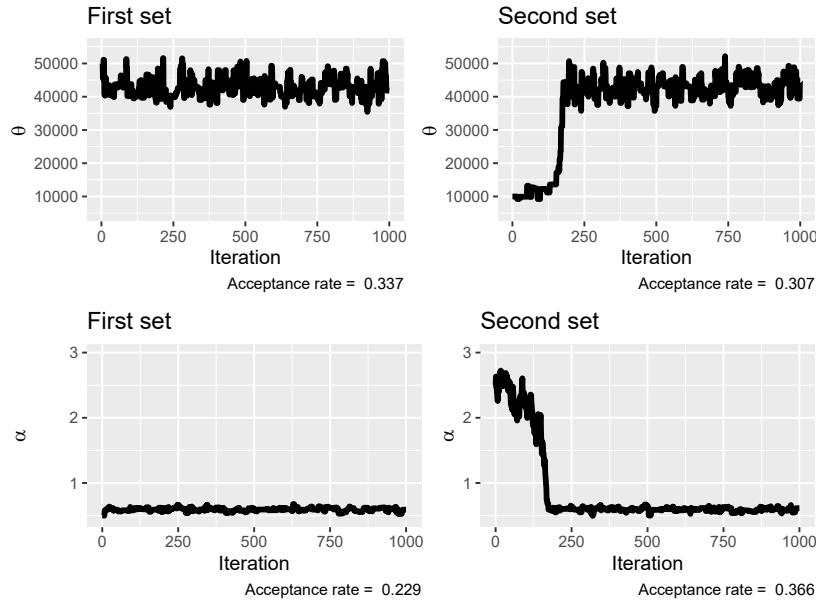
$$f_{\Theta, A}(\theta, \alpha) \propto 1, \quad \theta \in [0, \infty], \quad \alpha \in [0, \infty].$$

Find the posterior distribution of the parameter using 1,000 iterations of the Metropolis–Hastings sampler. Use a normal proposal with a proposal variance  $5 \times 10^{-7}$  for  $\theta$  and  $1 \times 10^{-2}$  for  $\alpha$ , and rely on  $\theta^{(0)} = 50,000$  and  $\alpha^{(0)} = 0.5$  to start the Metropolis–Hastings sampler. Redo the experiment with  $\theta^{(0)} = 10,000$  and  $\alpha^{(0)} = 2.5$ .

**Example Solution.** Starting from the the likelihood function and the prior distribution, we have that

$$h(\theta, \alpha) \propto \prod_{i=1}^N \frac{1}{\theta^\alpha \Gamma(\alpha)} x_i^{\alpha-1} e^{-\frac{x_i}{\theta}}.$$

You can learn more about the R code for this example at the online version of this book, [Actuarial Community \(2025\)](#).



**FIGURE 9.17: Trace plots based on two different starting parameter sets:  $\theta^{(0)} = 50,000$  and  $\alpha^{(0)} = 0.5$  (left panels), and  $\theta^{(0)} = 10,000$  and  $\alpha^{(0)} = 2.5$  (right panels)**

Clearly, from Figure 9.17, the initial parameter value matters: for the first set, the starting value is close to the posterior mode, meaning that the final sample does not depend much on the starting value. For the second set, on the other hand, it takes about 200 iterations to get closer to where most of the density resides. Having a burn-in in the case of Metropolis–Hastings sampler is therefore a good idea to reduce the impact of initial guesses on the final posterior distribution.

In the next subsection, we learn a few methods and metrics to diagnose the convergence of the Markov chains generated via MCMC methods.

#### 9.4.4 Markov Chain Diagnostics

There are many different tuning parameters in MCMC schemes, and they all have an impact on the convergence of the Markov chains generated by these methods. To understand the impact of these choices on the chains (e.g., number of iterations, length of burn-in, proposal distribution), we introduce a few methods to analyze their convergence.

### Examining Trace Plots and Autocorrelation

**Trace Plot.** The most elementary tool to assess whether MCMC chains have converged to the posterior distribution is the trace plot. As mentioned above, a trace plot displays the sequence of samples as a function of the iteration number, with the sample value on the  $y$ -axis and the iteration number on the  $x$ -axis. If the chain has converged, the trace plot should show a stable sequence of samples around the true posterior distribution that looks like a hairy caterpillar. However, if the chain has not yet converged, the trace plot may show a sequence of samples that still appear to be changing or have not yet settled into a stable pattern.

In addition to assessing convergence, trace plots can also be used to diagnose potential problems with MCMC algorithms, such as poor mixing or autocorrelation. For example, if the trace plot shows long periods of no change followed by abrupt jumps, this may indicate poor mixing and suggest that the MCMC algorithm needs to be adjusted or a different method should be used.

**Lag-1 Autocorrelation.** Another quantity that might be helpful is the lag-1 autocorrelation—the correlation between consecutive samples in a given chain:

$$\text{Cov} [\theta_i^{(m)}, \theta_i^{(m-1)}].$$

Note that if the autocorrelation is too high, it can indicate that the chain is not mixing well and is not sampling the posterior distribution effectively. This can result in poor convergence, longer run times, and decreased precision of the estimates obtained from the MCMC algorithm.

In addition to examining trace plots and computing autocorrelation coefficients, we can use other, more formal tools to evaluate whether the chains obtained are reliable and have converged.

### Comparing Parallel Chains

**Gelman–Rubin Statistic** Another way to assess convergence is to run multiple chains in parallel from different starting points and check if their behavior is similar. In addition to comparing their trace plots, the chains can be compared by using a statistical test—the Gelman–Rubin test of [Gelman and Rubin \(1992\)](#). The latter test compares the within-chain variance to the between-chain variance; to calculate the statistic, we need to generate a small number of chains (say,  $R$ ), each for  $M - M^*$  post-burn-in iterations.

If the chains have converged, the within-chain variance should be similar to the between-chain variance. Assuming the parameter of interest is  $\theta_i$ , the within-

chain variance is

$$W = \frac{1}{R(M - M^* - 1)} \sum_{r=1}^R \sum_{m=M^*+1}^M (\theta_{i,r}^{(m)} - \bar{\theta}_{i,r})^2,$$

where  $\theta_{i,r}^{(m)}$  is the  $m^{\text{th}}$  draw of  $\theta_i$  in the  $r^{\text{th}}$  chain and  $\bar{\theta}_{i,r}$  is the sample mean of  $\theta_i$  for the  $r^{\text{th}}$  chain. The between-chain variance is given by

$$B = \frac{M - M^*}{R - 1} \sum_{r=1}^R (\bar{\theta}_{i,r} - \bar{\theta}_i),$$

where  $\bar{\theta}_i$  is the overall sample mean of  $\theta_i$  from all chains. The Gelman–Rubin statistic is

$$\sqrt{\left( \frac{M - M^* - 1}{M - M^*} + \frac{R + 1}{R(M - M^*)} \frac{B}{W} \right) \frac{df}{df - 2}},$$

where  $df$  is the degrees of freedom from Student's  $t$ -distribution that approximates the posterior distribution. The statistic should produce a value close to 1 if the chain has converged. On the other hand, if the statistic value is greater than 1.1 or 1.2, this indicates that the chains may not have converged, and further analysis may be needed to determine why the chains are not mixing well.

### Calculating Effective Sample Sizes

**Effective Sample Size.** The effective sample size (ESS) is a measure of the number of independent samples obtained from an MCMC chain. Recall that in an MCMC chain, each sample is correlated with the previous sample; as a result, the effective number of independent samples is usually much smaller than the total number of samples generated by the MCMC algorithm. The ESS takes this correlation into account and provides an estimate of the number of independent samples that are equivalent to the correlated samples in the chain.

In general, a higher effective sample size indicates that the MCMC algorithm has produced more independent samples and is more likely to have accurately sampled the posterior distribution. A lower effective sample size, on the other hand, suggests that the MCMC algorithm may require further tuning or optimization to produce reliable posterior estimates.

The function `multiESS` of the R package `mcmcse` contains a function that gives the ESS of a multivariate Markov chain as described in [Vats et al. \(2019\)](#). The package also includes an estimate of the minimum ESS required for a specified relative tolerance level (see function `minESS`).

We now apply these various diagnostics to an example.

**Example 9.4.4. Markov Chain Diagnostics.** Consider the setup of Example 9.4.2. Using chains of 51,000 iterations and a burn-in of 1,000 iterations, calculate the various Markov chain diagnostics mentioned above.

You can learn more about the R code for this example at the online version of this book, [Actuarial Community \(2025\)](#).

**Example Solution.** Let us begin by generating five chains.

Figure 9.18 reports the trace plot for the first chain: it indeed looks like a hairy caterpillar, which is a good sign.

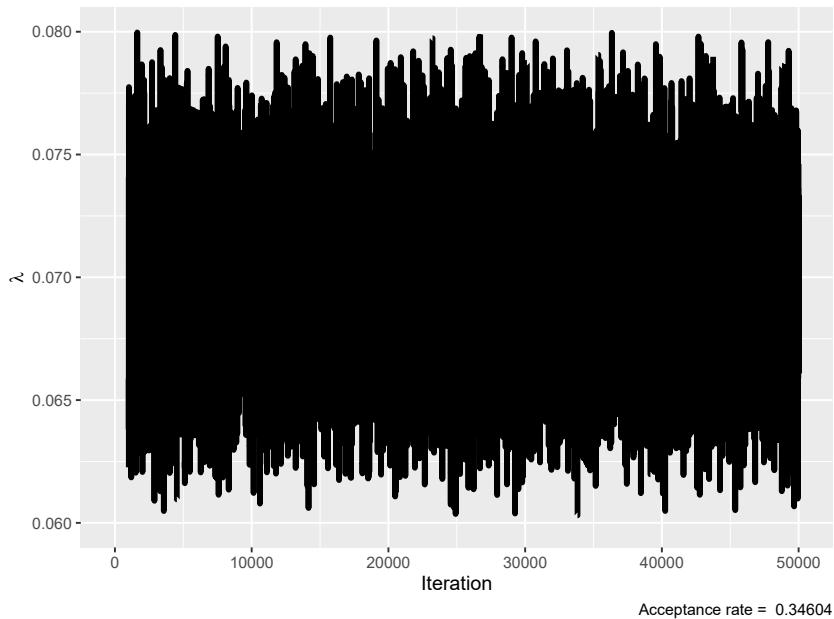


FIGURE 9.18: Trace plot for parameter  $\lambda$

The lag-1 autocorrelation coefficient is 0.651510865

The autocorrelation is also mild at 65%, again pointing towards good convergence behavior.

The Gelman–Rubin statistic is 1.00012696

The Gelman–Rubin statistic is very close to 1 in this case, meaning that the chains converged.

The ESS is 9927.29934 and the minimum ESS is 6146

The last diagnostic refers to the ESS, and its comparison to the minimum ESS. In our case, the ESS is about 9,927, and the minimum ESS is 6,146.

Since our ESS is above the minimum, we know we have a large enough sample to adequately capture the posterior distribution of  $\lambda$ .

---



---

## 9.5 Bayesian Statistics in Practice

---

In Section 9.5, you learn how to:

- Describe the main computing resources available for Bayesian statistics and modeling.
  - Apply one of them to loss data.
- 

Fortunately for end users, some of these methods are readily available in R, meaning that they are quite accessible. Some popular computing resources used in Bayesian statistics are listed below:

- **RSTAN**, named in honor of Stanislaw Ulam, is an R implementation of the widely used STAN probabilistic programming language for Bayesian statistical modeling and inference. It is highly flexible and allows users to define complex statistical models.
- **nimble** stands for Numerical Inference for Bayesian and Likelihood Estimation and is an R package designed for statistical computing and hierarchical modeling. **nimble** provides a high-level programming language that allows users to define complex statistical models with ease.
- **R2OpenBUGS** allows R users to use OpenBUGS, a classic and widely-used software package for Bayesian data analysis. It uses MCMC techniques like Gibbs sampling to obtain samples from the posterior distribution.
- **rjags** is an R implementation of the JAGS (Just Another Gibbs Sampler) program. It is an open-source software that was developed as an extension of BUGS. It provides a platform-independent engine for the BUGS language, allowing for the use of BUGS models in various environments. Like BUGS, JAGS is also used for Bayesian analysis through MCMC sampling techniques.

In what follows, we will use the **nimble** package in the context of loss data.

---

**Example 9.5.1. The **nimble** package.** Similar to the setup of Example 9.4.2, consider that each policyholder's claim count (frequency) is distributed

as a Poisson random variable such that

$$p_{N_i | \Lambda=\lambda}(n_i) = \frac{\lambda^{n_i} e^{-\lambda}}{n_i!},$$

where  $n_i$  is the number of claims associated with the  $i^{\text{th}}$  policyholder. Unlike the previous example, however, let us assume an inverse gamma prior with a shape parameter of 2 and a scale parameter of 5.<sup>16</sup>

Find the posterior distribution of the parameter by creating a chain of 51,000 iterations and a burn-in of 1,000 iterations using the **nimble** package.

**Example Solution.** First, we need to define the model using the ‘nimble’ language. Simply put, the model is comprised of a likelihood density and a prior density. The former links the observations to a Poisson distribution with parameter  $\lambda$ , and the latter states the prior distribution, which is inverse gamma with shape and scale parameters of 2 and 5, respectively.

```
claimmodel <- nimbleCode({
  for (i in 1:N) {
    # Likelihood
    count[i] ~ dpois(lambda)
  }
  # Prior distribution
  lambda ~ dinvgamma(shape = 2, scale = 5)
})
```

Then, we define the data, the constant (i.e., the number of observations in this case), the parameter list (i.e., only  $\lambda$  here), and the initial value set to 0.05 in this illustration.

```
claimdata <- list(count = sgautonb$Clm_Count)
claimconstant <- list(N = length(sgautonb$Clm_Count))
claimparameters <- c("lambda")
claiminitial <- list(lambda = 0.05)
```

The MCMC chain is then run using for 51,000 iterations and a burn-in of 1,000 iterations.

```
mcmcoutput <- nimbleMCMC(code = claimmodel, data = claimdata, constants = claimconstant,
  inits = claiminitial, monitors = claimparameters, niter = 51000, nburnin = 1000,
  nchains = 1)
save(mcmcoutput, file = "../IntermediateCalcs/BayesChap/Example951.Rdata")
```

---

<sup>16</sup>Note that the inverse gamma prior combined with a Poisson distribution does not generally lead to closed-form posterior densities and thus requires us to use MCMC methods.

Finally, we display the trace plot, obtain the histogram of the posterior distribution of  $\lambda$ , and compute some descriptive statistics of the parameter.

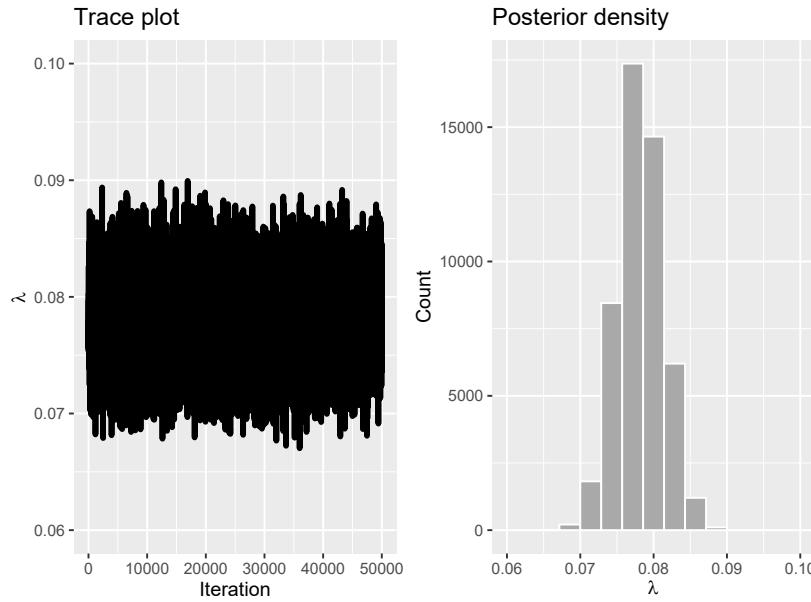


FIGURE 9.19: Trace plot and posterior density for parameter  $\lambda$

The posterior mean of the parameter is 0.0781913941

The posterior standard deviation of the parameter is 0.00309462947

This simple illustration demonstrates the simplicity of utilizing R packages to generate MCMC chains, all without the need for writing extensive code. For more details on the `nimble` package, see [Valpine et al. \(2017\)](#).

## 9.6 Further Resources and Contributors

Many great books exist on Bayesian statistics and MCMC schemes. We refer the interested reader to [Bernardo and Smith \(2009\)](#) and [Robert and Casella \(1999\)](#) for an advanced treatment of these topics.

A number of academic articles in actuarial science relied on Bayesian statistics and MCMC schemes over the past 40 years; see, for instance, [Heckman and Meyers \(1983\)](#), [Meyers and Schenker \(1983\)](#), [Cairns \(2000\)](#), [Cairns et al. \(2006\)](#), [Hartman and Heaton \(2011\)](#), [Bermúdez and Karlis \(2011\)](#), [Hartman and Groendyke \(2013\)](#), [Fellingham et al. \(2015\)](#), [Bignozzi and Tsanakas](#)

(2016), Bégin (2019), Huang and Meng (2020), Bégin (2021), Cheung et al. (2021), and Bégin (2023).

### Contributors

- **Jean-François Bégin**, Simon Fraser University, is the principal author of the initial version of this chapter. Email: [jbegin@sfu.ca](mailto:jbegin@sfu.ca) for chapter comments and suggested improvements.
- Chapter reviewers include: Brian Hartman, Chun Yong, Margie Rosenberg, and Gary Dean.



# 10

---

## Premium Foundations

---

*Chapter Preview.* Setting prices for insurance products, i.e., premiums, is an important task for actuaries and other data analysts. This chapter introduces the foundations for pricing non-life products.

The presentation of this chapter follows the premium equation.

- In Section 10.2, we first present the sources of information that support premium development.
  - We discuss this development of the pure premiums in Section 10.5.
  - In Section 10.6, we discuss fixed and variable non-claim expenses.
  - In Section 10.7.3, we discuss the provision for profit.
  - Section 10.9 summarizes alternative premium principles that incorporate uncertainty into our pricing.
- 

### 10.1 Introduction to Ratemaking

---

In this section, you will learn how to:

- Describe relationship between exposures, rates, and premiums
  - Describe the components of the rate
- 

This chapter explains how you can determine the appropriate price for an insurance product. As described in Section 1.2, one of the core actuarial functions is ratemaking, where the analyst seeks to determine the right price for a risk.

A price is the consideration exchanged for a good or service. In insurance, we refer to this consideration as the premium and the service provided by the insurer is protection against contingent events.

The amount of protection will vary by risk being insured. For example, in homeowners insurance, the amount of insurance protection depends on the

home value. In life insurance, the amount of protection may depend on a policyholder's financial status (e.g., income and wealth) and their perceived need for financial security. So, it is common to express insurance prices as a unit of the protection being purchased, for example, a price per thousand dollars of coverage on a home or benefit in the event of death. We refer to the unit of protection as the exposure. These prices/premiums are known as rates because they are expressed in standardized units.

Unlike other products, the costs of insurance protection are not known at the sale of the contract. If the insured contingent event, such as an automobile accident or the loss of life, does not occur, then the contract costs are only administrative (e.g., to set up the contract) and are relatively minor. If an insured event occurs, then the cost includes not only administrative costs but also claim payment(s) and expenses to settle claims. So, the cost is random when the contract is written, and protection from that randomness is the basis of insurance.

Because costs are unknown at the time of sale, insurance pricing differs from common economic approaches. This chapter introduces traditional actuarial approaches to determine prices as a function of insurance costs. Insurance involves a promise of the insurer to pay a claim when presented by the insured. For this reason, insurance is a regulated business, particularly for personal lines insurance. The role of the regulator is to ensure that the insurer is able to satisfy its promise to its policyholders. In executing this mandate, the regulator often requires the insurer to file support for its rates. The regulator will review that filing to determine whether those rates are reasonable, not excessive, not inadequate, and not unfairly discriminatory.

The actuarial pricing approach we present is sufficient for some insurance markets, such as personal automobile or homeowners, where the insurer has a portfolio of many similar independent risks. However, there are other insurance markets where actuarial prices only provide an input to general market prices. To reinforce this distinction, actuarial cost-based premiums are sometimes known as technical prices.

To develop technical prices, it is helpful to think of a premium as revenue source that provides for

- Pure Premium - Claim payments are amounts due to the insured under the terms of the insurance contract. Pure premiums include claim payments costs to administer and investigate such claims.
- Insurer expenses - *Non-claim Expenses* include insurer costs that vary by premium (such as sales commissions), and those that do not (such as build-

ing costs and employee salaries). We include those costs through the Fixed Expenses and Variable Expense Rate of the (10.2).

- Profit - An insurer requires capital to support operations. The capital provider will reasonably expect to earn a profit from insuring risk. Insurers have two sources of profit: underwriting income and investment income.

We formalize this relationship in our simplified premium equation.

$$\text{Premium} = \frac{\text{Pure Premiums} + \text{Fixed Expenses}}{1 - \text{Variable Expense Rate} - \text{Profit}} \quad (10.1)$$

where

$$\text{Pure Premiums} = \frac{\text{Estimated Claims and Claims Adjustment Expense}}{\text{Exposures}}$$

or

$$\text{Pure Premiums} = \frac{\text{Estimated Claim Counts}}{\text{Exposures}} \times \frac{\text{Estimated Claims and Claims Adjustment Expense}}{\text{Estimated Claim Counts}}$$

This simplified premium equation promotes a general understanding of the Relationship of insurance costs and revenue. We refer to this equation as the simplified premium equation because (i) it does not include explicit consideration for investment income, and (ii) we combine consideration of claims and claims adjustment expenses. We will refine the simplified premium equation to consider these items later in this chapter.

We observe that the *pure premium* in equation (10.1) is a ratio of claims and exposures. We discuss the development of the claims provision in Section 10.3 and the development of exposures in Section 10.4.

## 10.2 Data Sources

In this section, you will learn how to:

- Describe the types of data used to develop rates

Insurers consider aggregate information for ratemaking such as exposures, premiums, expenses, claims, and payments. This aggregate information is also useful for managing an insurer's activities. The information is typically summarized in financial reports which are commonly compiled at least annually

and often quarterly. At any given financial reporting date, information about recent policies and claims will be ongoing and necessarily incomplete; this section introduces concepts for projecting risk information so that it is useful for ratemaking purposes.

Insurers generally store information about insured risks, such as exposures, premiums, claim counts, losses, and rating factors, in a relational database that will include:

- *policy database* - contains information about the risk being insured, the policyholder, and the contract provisions
- *claims database* - contains information about each claim. The claims database is linked to the policy database.
- *payment database* - contains information on each claims transaction, typically payments but may also include changes to *case reserves*. The payment database is linked to the claims database.

Insurers will aggregate the information in these detailed databases to develop the information needed for financial reports. As described in this chapter, insurers' actuaries will also use this information to develop the premiums.

---

### 10.3 Claims

---

In this section, you will learn how to:

- Describe the basis for the provision for claims, the numerator of the pure premium
  - Adjust claims to the level of the prospective period
- 

The terms loss and claim refer to the amount of compensation paid or payable to the claimant under the terms of the insurance policy. Definitions can vary:

- Sometimes, *claim* is used interchangeably with the term *loss*.
- In some insurance and actuarial sources, *loss* is the amount of damage sustained in an insured event. The *claim* is the amount paid by the insurer. Differences between *loss* and *claim* amounts are typically due to the coverage terms such as deductibles and policy limits.
- In economics, a *claim* is a demand for payment by an insured or by an injured third party under the terms and conditions of the insurance contract, and the *loss* is the amount paid by the insurer.

This text will follow the convention of the second bullet.

Also, there are two categories of claim adjustment expenses.

- Allocated claim adjustment expenses are attributed to a specific claim and are generally comprised of investigation and legal expenses to defend or settle the claim.

Claims and allocated claim adjustment expenses are sometimes inversely correlated, as additional defense expenses may result in lower claim payments. In this section, references to claims also include allocated claims adjustment expenses.

- Unallocated claim adjustment expenses cannot be assigned to individual claims (e.g., claim adjuster salaries.) Actuaries often review claims and allocated claim adjustment expenses in aggregate, as the latter is often a function of the former.

Insurers will include a provision for unallocated claims adjustment expenses either as a percentage of claims and allocated claims expenses or premiums, or a combination thereof. We discuss unallocated claims adjustment expenses separately in Section 10.3.2.

### 10.3.1 Estimated Ultimate Claims

Recall that a claim is the amount paid or payable to claimants under the terms of insurance policies. In more detail, one can consider *paid claims*, those losses for a particular period that have actually been paid to claimants. When there is an expectation that payment will be made in the future, a claim will have an associated *case reserve* representing the estimated amount of that payment. Case adjusters establish case reserves separately for each open claim based on the information available. In addition, *reported claims*, also known as *case incurred claims* or *incurred claims*, are the sum of paid claims and case reserves. The *ultimate claim* is the amount required to close and settle all claims for a defined group of policies. We describe the estimation of ultimate claims and claims adjustment expenses in Section ??.

Alternatively, we can estimate projected claims and claims expenses as the product of projected claim frequency and claims severity:

$$\text{Claims and Claims Expenses}_i^{(t)} = E[X] \times E[N]$$

where:

- $E[X]$  is the projected average ultimate severity per claim, and

- $E[N]$  is the projected ultimate number of claims.

We note the frequency-severity alternatives to support our discussion of trends in the following section.

### 10.3.2 Adjustments to Claims and Allocated Claims Adjustment Expenses

In this section, we review adjustments to experience period ultimate claims that are required to support the development of prospective rates. These adjustments include trending, large loss adjustments and provisions for catastrophes. Finally, we discuss approaches to incorporate unallocated claims adjustment expense.

#### Trending

Each of the years of the experience period has a different underlying cost level; our goal is to estimate claims at the cost level of the prospective policy period. Consider, for example, if costs were rising at a rate of 5% per annum. All else equal, the estimated ultimate cost for time  $t+2$  would be  $1.05^2$  times the costs of claims from time  $t$ . Trending is the process of adjusting ultimate losses from the cost level of the experience period to prospective cost levels.

Actuaries will often consider separate trends for the frequency of claims and the severity of claims. Actuaries often state past trends separately from future trends. Past trends reflect changes that have taken place between the experience period of the rate calculation and the valuation date. Future trends reflect expectations of change between the valuation date and the prospective policy period.

There are various approaches to estimating severity trend rates. Two common approaches include the estimation of trend rates based on external cost indices and the estimation of trend rates based on claims experience. The former approach generally uses government data such as the consumer price index or components thereof. In the latter approach, actuaries will often fit regression models to discern the rate of change in average claims values over time.

Due to the lack of external indices that would be appropriate as a basis for claims frequency models, actuaries generally either estimate frequency trend based on company experience or assume that the frequency trend is 0%.

It is also common to review pure premium trends directly. Although the pure premium trend is effectively a combination of the frequency and severity trends, direct analysis of pure premiums may mask underlying changes in frequency and severity when they are inversely correlated.

### Large Loss and Catastrophe Provisions

Consider, for example, if a five-year experience period included a one-in-20-year event. If we did not adjust the data, we would effectively overestimate claim amounts for that category of claims.

In ratemaking, we remove these unusual large losses and catastrophe losses from the experience period data, and then add a provision consistent with the longer-term average cost of large losses or catastrophes. Although large loss adjustments are commonly based on the insurer's claims experience, the provision for catastrophes is often based on models developed by specialists. Adjustments for catastrophes are more common in property insurance, while adjustments for large losses are more common in liability insurance.

### Unallocated Claims Adjustment Expenses

Some insurers include unallocated claims adjustment expenses as a percentage of claims and allocated claims adjustment expenses, while other insurers include unallocated claims adjustment expenses as a percentage of premiums. In our discussion, we use the former approach. For insurers that use the latter approach, the inclusion of a provision for unallocated claims adjustment expenses would follow that described below for other non-claim expenses. Generally, insurers estimate  $UE$  by reviewing historical ratios of those payments to claims and allocated claims adjustment expense payments.

---

## 10.4 Exposures

---

In this section, you will learn how to:

- Describe the consideration exposures in the developing pure premiums
- Select an exposure base
- Adjust historical exposures to the level of the prospective period

---

The denominator of the pure premium equation is “exposure.” We use exposures to standardize heterogeneous risks. To explain exposures, we can consider *scale distributions* that we learned about in Chapter 4. To recall a scale distribution, suppose that  $X$  has a parametric distribution and define a rescaled version  $R = X/E$ ,  $E > 0$ . If  $R$  is in the same parametric family as  $X$ , then the distribution is said to be a scale distribution. As we have seen, the gamma, exponential, and Pareto distributions are examples of scale distributions.

Intuitively, the idea behind exposures is to make risks more comparable to one another. For example, it may be that risks  $X_1, \dots, X_n$  are from different distributions and yet, with the choice of the right exposures, the rates  $R_1, \dots, R_n$  are from the same distribution. Here, we interpret the rate  $R_i = C_i/E_i$  as the loss divided by exposure.

**Table 10.5.1** provides a few examples.

**Table 10.5.1. Commonly used Exposures in Different Types of Insurance**

Type of Insurance	Exposure Basis
Personal Automobile	Earned Car Year, Amount of Insurance Coverage
Homeowners	Earned House Year, Amount of Insurance Coverage
Workers Compensation	Payroll
Commercial General Liability	Sales Revenue, Payroll, Square Footage, Number of Units
Commercial Business Property	Amount of Insurance Coverage
Physician's Professional Liability	Number of Physician Years
Professional Liability	Number of Professionals (e.g., Lawyers or Accountants)
Personal Articles Floater	Value of Item

#### 10.4.1 Criteria for Choosing an Exposure

An exposure base should meet the following criteria. It should:

- be an accurate measure of the quantitative exposure to loss
- be easy for the insurer to determine (at the time the policy is initiated) and not subject to manipulation by the insured,
- be easy to understand by the insured and easy to calculate by the insurer,
- consider any preexisting exposure base established within the industry.

To illustrate, consider personal automobile coverage. Instead of the exposure basis “earned car year,” a more accurate measure of the quantitative exposure to loss might be number of miles driven. Historically, this measure had been difficult to determine at the time the policy is issued and subject to potential manipulation by the insured, so it was not typically used. Modern telematic devices that allow for accurate mileage recording support the use of this exposure base in some marketplaces.

As another example, the exposure measure in commercial business property, e.g., fire insurance, is typically the amount of insurance coverage. As property values grow with inflation, so will the amount of insurance coverage. Thus, rates quoted on a per amount of insurance coverage are less sensitive to inflation.

### 10.4.2 Written and Earned Exposures

In developing premiums and rates, it's important that we use claims information and exposure information that is comparable. Most ratemaking uses an accident year approach. In this approach, we relate claims incurred during a specified period to the premium or exposure "earned" during that same period without consideration of the period in which the underlying policy was written. For example, a 12-month policy issued on 1 July 2019 insures claims events in 2019 or 2020, and the claims are assigned to the year of the event. Generally, we earn premiums and exposures on a pro-rata as to time basis as presented in **Table 10.5.2**, which displays illustrative calculations for a portfolio of four illustrative policies.

Table 10.5.2. Exposures for Four 12-Month Policies

<i>Policy</i>	Effective Date	Written 2019	Exposure 2020	Earned 2019	Exposure 2020	Unearned 1/1/2019	Exposure 1/1/2020	In-Force Exposure 1/1/2020
A	1 Jan 2019	1.00	0.00	1.00	0.00	0.00	0.00	0.00
B	1 April 2019	1.00	0.00	0.75	0.25	0.25	0.00	1.00
C	1 July 2019	1.00	0.00	0.50	0.50	0.50	0.00	1.00
D	1 Oct 2019	1.00	0.00	0.25	0.75	0.75	0.00	1.00
<i>Total</i>		4.00	0.00	2.50	1.50	1.50	0.00	3.00

### 10.4.3 Adjustments to Exposures

#### Exposure Trend

Sometimes exposure units are inflation sensitive. For example, payroll is a common exposure base for workers compensation coverage. Even if the insured firm does not grow, its payroll may increase due to wage inflation. We refer to the adjustment applied to inflation sensitive exposures as exposure trend.

## 10.5 Pure Premiums

In this section, you will learn how to:

- Calculate the expected pure premium

The *pure premium* in equation (10.1) is a random variable, and so, as a baseline, we use the *expected costs* to determine rates. To develop our initial understanding, we will consider the insurer that enters into many contracts with risks that are similar except, by pure chance, in some cases, there are losses on some contracts but not on others. The insurer is obligated to pay the total amount of claim payments for all contracts. If the risks are similar, then all policyholders are equally likely to contribute to the total loss. From probability theory, specifically the law of large numbers, we know that the average of iid risks is close to the expected amount, so we use the expectation as a baseline pricing principle.

In this chapter, we present the development of average premium levels for a portfolio of homogeneous risks. In Chapter ??, we present approaches to develop classification plans which adjust those average premiums to recognize various risk characteristics. In Chapter ??, we present approaches to develop premiums that consider the claim experience of an individual insured.

### 10.5.1 Experience Period

To develop expected pure premiums, actuaries will typically review claims and exposure experience over a multi-year (typically three to seven years) period. The use of a multi-year period smooths the year-to-year randomness. We refer to this multi-year period as the experience period.

### 10.5.2 Expected Pure Premium

The expected pure premium is generally calculated as the weighted average of the observations in the experience period. The weights balance responsiveness to more recent experience and the stability of a longer-term average.

$$\text{Pure Premium}^{(t)} = \text{Exposure}_t \times \sum_{i=1}^n w_i \frac{\text{Ultimate Claims and Claims Expenses}_i^{(t)}}{\text{Exposure}_i^{(t)}}$$

where:

- $w_i$  is the weight for year  $i$  in an  $n$  year experience period.

The superscript ( $t$ ) indicates that the ultimate claim estimate for accident year  $i$  is adjusted to the level of the prospective program period  $t$ . We discussed these adjustments in Section 10.3.2. The following equation demonstrates this process of adjustment.

$$\begin{aligned} \frac{\text{Claims and Claims Expenses}_i^{(t)}}{\text{Exposure}_i^{(t)}} &= C_i^{\text{xLL}, \text{xCat}} \times T_i^{(t)} \times LL \times CP \times UE \\ &= \text{Exposure}_i \times E_i^{(t)} \end{aligned}$$

where:

- $C_i^{xLL,xCat}$  is the estimated ultimate claims for year  $i$ , excluding large losses and catastrophes
- $T_i$  is a claim trend factor to adjust year  $i$  experience to the cost level of year  $t$
- $E_i^{(t)}$  is an exposure trend factor to adjust year  $i$  experience to the cost level of year  $t$
- $LL$  is a large loss factor
- $CP$  is a catastrophe provision
- $UE$  is the unallocated claims adjustment expense factor

We discussed these adjustments earlier in this chapter.

---

## 10.6 Non-Claim Expenses

---

In this section, you will learn how to:

- Describe the consideration of operational expenses in the development of premiums
- 

Non-claim insurer Operating expense costs include commissions, premium taxes, and other expenses such as salaries, rent, and inspections.

- Some expenses (such as commissions and premium taxes) vary with premiums are “variable” or “premium variable expenses.”
- Other expenses (such as general administrative and head office costs) are not proportional to the premium.

For non-claim expenses, insurers will typically rely on either historical expense ratios, budgeted amounts, or financial forecasts.

We include fixed expenses in our premium equation on a per-exposure basis and we include variable expenses as a rate per unit premium.

---

## 10.7 Investment Income

---

In this section, you will learn how to:

- Describe the consideration of the timing of cash flows in the development of the rate
  - Calculate a required provision for underwriting profit
- 

A portion of the required profit is earned from investment income from two sources: policyholder cash flows and investment of the insurer's surplus. To the extent that investment income is insufficient to provide the required rate of return, the premiums will also need to include an underwriting profit provision.

As we described, we presented a simplified premium equation in Section 10.5.2 to promote the understanding of the claims and expense provisions in Section 10.3 and exposures in Section 10.6. We now refine the equation to consider investment income. We now consider the other source of an insurer's profit, investment income.

#### 10.7.1 Investment Income on Policyholder Cash Flows

We first consider policyholder cash flows, i.e., premiums, claims, claim adjustment expenses, and non-claim expenses. We consider investment income on policyholder cash flows by discounting each of the cash flows of each of these components of the premium equation.

- There may be a delay in the insurer's receipt of premium, perhaps because the insurer offers payment plans to the insured.
- Claims and claims adjustment expenses are paid over a period that typically extends beyond the policy term. Generally, property coverages have the shortest payment stream, with all claims being settled and paid over a period that extends between 2 and 5 years, depending on the complexity of the determination of damages. Litigated liability coverages will have intermediate payment streams that range from three to 10 years. Finally, coverages such as workers compensation, offer lifetime benefits that can extend forty years or longer.
- Non-claim expenses are generally paid over the term of the policy period.

We can rewrite our premium equation to capture the discounting. We replace the unity in the denominator with a premium delay factor and we discount the claims and claims adjustments expenses (in the pure premium) and non-claim expenses.

$$\text{Premium} = \frac{\text{Discounted Pure Premiums} + \text{Discounted Fixed Expenses}}{\text{Premium Delay} - \text{Variable Expense Rate} - \text{Profit}}.$$

We recognize that the discounting effect on the numerator is significantly

greater than the effect on the denominator. As a result, consideration of investment income on policyholder cash flows serves to reduce the premium.

The consideration of profit serves in the denominator serves to increase the required premium. We now turn to the determination of that profit provision.

### 10.7.2 Investment Income on Surplus

The insurer's surplus is also comprised of invested assets which provide a rate of return. In ratemaking we assume that the investment income on surplus is earned over the policy term, generally 12 months. Investment income of surplus will reduce the required profit provision. For example, if the insurer were able to earn a rate of return on assets of 5% per annum, then the insurer would realize a return of 2.5% of premiums assuming a 2:1 premium: surplus ratio.

### 10.7.3 The Underwriting Profit Provisions

An insurer requires capital to support operations. The insured pays a premium for the promise of the insurer to pay a claim in the future. Capital serves as protection for the policyholder in the event that premiums are insufficient to pay claims. The capital provider will reasonably expect to earn a profit to insure the risk and subject its capital to loss. Generally, the required profit is expressed as after-tax return on equity.

If the profit provision in the premium equation were 0, then the premium would equal the present value of the present value of cash flows of the insurance policy. However, as we discussed the insurer will require a return on its capital. Generally, coverages that are riskier, i.e., have more variability, will require more supporting capital. Every claim submitted to the insurer has access to all of the capital of the insurer. In insurance, capital is often referred to as surplus. For ratemaking, we notionally allocate capital to coverage using premium to surplus ratios and we state the required rate of return on an after-tax basis. We have to convert that return to a "percent of premium basis" to include in our premium equation. For example, if we assume a 2:1 premium:surplus ratio, a required after-tax rate of return of 12% and a tax rate of 30%, then the profit provision in the premium would be:

$$\begin{aligned} \frac{12\% \text{ after tax return}}{\text{surplus}} &\times \frac{1 \times \text{surplus}}{2 \times \text{premium}} \times \frac{1 \text{ pre-tax}}{0.7 \text{ after tax}} \\ &= \frac{8.6\% \text{ pre-tax return}}{\text{premium}}. \end{aligned}$$

We can then reduce the required underwriting profit to consider investment

income on surplus. Using the example of Section 10.7.2, the resulting required underwriting profit provision would reduce from 8.6% to 6.1%.

---

## 10.8 The Premium Equation

---

In this section, you learn how to:

- Calculate the rate for a class of risk
  - Calculate premiums
- 

We can now remove the simplifying assumptions included in Equation (10.1) and provide our final premium equation. The term “pure premium” can be used to refer to rate per exposure unit of provision for claims costs included in the premium for an insured (which may have a quantum of exposure more or less than one exposure unit). In this section, we use the latter definition.

$$\text{Premium} = \frac{\text{Discounted Pure Premium} + \text{Discounted Fixed Expenses}}{\text{Premium Delay} - \text{Variable Expense Rate} - \text{Required Underwriting Profit}}. \quad (10.2)$$

---

## 10.9 Pricing Principles

---

In this section, you learn how to:

- Describe common actuarial pricing principles
  - Describe properties of pricing principles
  - Choose a pricing principle based on a desired property
- 

Approaches to pricing vary by the type of contract. For example, personal automobile is a widely available product throughout the world and is known as part of the *retail general insurance* market in the United Kingdom. Here, one can expect to do pricing based on a large pool of independent contracts, a situation in which expectations of losses provide an excellent starting point. In contrast, an actuary may wish to price an insurance contract issued to a large employer that covers complex health benefits for thousands of employees.

In this example, knowledge of the entire distribution of potential losses, not just the expected value, is critical for starting the pricing negotiations. To cover a range of potential applications, this section describes general premium principles and their properties that one can use to decide whether or not a specific principle is applicable in a given situation.

### 10.9.1 Premium Principles

The prior sections of this chapter introduce traditional actuarial pricing principles that provide a price based only target rates of return and the cost to insure the risk; the price does not depend on the demand for insurance.

Assume that the loss  $X$  has distribution function  $F(\cdot)$  and that there exists some rule (which in mathematics is known as a *functional*), say  $H$ , that takes  $F(\cdot)$  into the positive real line, denoted as  $P = H(F)$ . For notation purposes, it is often convenient to substitute the random variable  $X$  for its distribution function and write  $P = H(X)$ . **Table 10.8.1** provides several examples.

Table 10.8.1. Common Premium Principles

Description	Definition ( $H(X)$ )
Net (pure) premium	$E[X]$
Expected value	$(1 + \alpha)E[X]$
Standard deviation	$E[X] + \alpha SD(X)$
Variance	$E[X] + \alpha Var(X)$
Zero utility	solution of $u(w) = E[u(w + P - X)]$
Exponential	$\frac{1}{\alpha} \log E[e^{\alpha X}]$

A premium principle is similar to a risk measure that is introduced in Section ???. Mathematically, both are rules that map the loss rv of interest to a numerical value. From a practical viewpoint, a premium principle provides a guide as to how much an insurer will charge for accepting a risk  $X$ . In contrast, a risk measure quantifies the level of uncertainty, or riskiness, that an insurer can use to decide on a capital level to be assured of remaining solvent.

The net, or pure, premium essentially assumes no uncertainty. The expected value, standard deviation, and variance principles each add an explicit loading for uncertainty through the risk parameter  $\alpha \geq 0$ . For the principle of zero utility, we think of an insurer with utility function  $u(\cdot)$  and wealth  $w$  as being indifferent to accepting and not accepting risk  $X$ . In this case,  $P$  is known as an indifference price or, in economics, a reservation price. With exponential utility, the principle of zero utility reduces to the exponential premium principle, that is, assuming  $u(x) = (1 - e^{-\alpha x})/\alpha$ .

For small values of the risk parameters, the variance principle is approximately equal to exponential premium principle, as illustrated in the following special case.

**Special Case: Gamma Distribution.** Consider a loss that is gamma distributed with parameters  $\eta$  and  $\theta$  (we usually use  $\alpha$  for the location parameter but, to distinguish it from the risk parameter, for this example we call it  $\eta$ ). From the Appendix Chapter ??, the mean is  $\eta \theta$  and the variance is  $\eta \theta^2$ . Using  $\alpha_{Var}$  for the risk parameter, the variance premium is  $H_{Var}(X) = \eta \theta + \alpha_{Var}(\eta \theta^2)$ . From this appendix, it is straightforward to derive the well-known moment generating function,  $M(t) = E[e^{tX}] = (1-t\theta)^{-\eta}$ . With this and a risk parameter  $\alpha_{Exp}$ , we may express the exponential premium as

$$H_{Exp}(X) = \frac{-\eta}{\alpha_{Exp}} \log(1 - \alpha_{Exp}\theta).$$

To see the relationship between  $H_{Exp}(X)$  and  $H_{Var}(X)$ , we choose  $\alpha_{Exp} = 2\alpha_{Var}$ . With an approximation from calculus ( $\log(1-x) = -x - x^2/2 - x^3/3 - \dots$ ), we write

$$\begin{aligned} H_{Exp}(X) &= \frac{-\eta}{\alpha_{Exp}} \log(1 - \alpha_{Exp}\theta) = \frac{-\eta}{\alpha_{Exp}} \left\{ -\alpha_{Exp}\theta - (\alpha_{Exp}\theta)^2/2 - \dots \right\} \\ &\approx \eta\theta + \frac{\alpha_{Exp}}{2}(\eta\theta^2) = H_{Var}(X). \end{aligned}$$

### 10.9.2 Properties of Premium Principles

Properties of premium principles help guide the selection of a premium principle in applications. Table 10.8.2 provides examples of properties of premium principles.

Table 10.8.2. Common Properties of Premium Principles

Description	Definition
Nonnegative loading	$H(X) \geq E[X]$
Additivity	$H(X_1 + X_2) = H(X_1) + H(X_2)$ , for independent $X_1, X_2$
Scale invariance	$H(cX) = cH(X)$ , for $c \geq 0$
Consistency	$H(c + X) = c + H(X)$
No rip-off	$H(X) \leq \max\{X\}$

This is simply a subset of the many properties quoted in the actuarial literature. For example, the review paper of Young (2014) lists 15 properties. See also the properties described as *coherent axioms* that we introduce for risk measures in Section ??.

Some of the properties listed in [Table 10.8.2](#) are mild in the sense that they will nearly always be satisfied. For example, the *no rip-off* property indicates that the premium charge will be smaller than the largest or “maximal” value of the loss  $X$  (here, we use the notation  $\max\{X\}$  for this maximal value which is defined as an “essential supremum” in mathematics). Other properties may not be so mild. For example, for a portfolio of independent risks, the actuary may want the *additivity* property to hold. It is easy to see that this property holds for the expected value, variance, and exponential premium principles but not for the standard deviation principle. Another example is the *consistency* property that does not hold for the expected value principle when the risk loading parameter  $\alpha$  is positive.

The *scale invariance* principle is known as *homogeneity of degree one* in economics. For example, it allows us to work in different currencies (e.g., from dollars to Euros) as well as a host of other applications. Although a generally accepted principle, we note that this principle does not hold for a large value of  $X$  that may border on a surplus constraint of an insurer; if an insurer has a large probability of becoming insolvent, then that insurer may not wish to use linear pricing. It is easy to check that this principle holds for the expected value and standard deviation principles, although not for the variance and exponential principles.

---

## 10.10 Reviewing Rate Adequacy

After establishing the initial premiums, insurance company actuaries will perform rate reviews to measure the current adequacy of those rates. For many regulated coverages (typically, personal lines insurance), actuaries file those rate reviews with the insurance regulator. Actuaries review rates regularly as rate levels require updates to keep pace with inflationary pressures. At times, the required rate will have a decreasing trend; for example with improvements in vehicle safety technology or workplace safety. Of course, the primary purpose of the rate is to test whether the experience of the rate program is consistent with loss and expense assumptions underlying the current rates.

### 10.10.1 The Loss Ratio Method

The “loss ratio method” is a common approach to assess rate adequacy. The loss ratio is the ratio of loss to the premium.

$$\text{Loss Ratio} = \frac{\text{Loss}}{\text{Premium}}.$$

When determining premiums, it is a bit counter-intuitive to emphasize this ratio because the premium component is built into the denominator. As we will see, the loss ratio method develops rate **changes** rather than rates; we can use rate changes to adjust the current rate to the current costs levels.

We calculate rate changes by comparing the loss ratio of the experience period to the target loss ratio. This adjustment factor is then applied to current rates to determine new indicated rates.

### 10.10.2 Target Loss Ratio

Let us return to equation (10.2). Noting that the “pure premium” is the provision for loss in the rates, we can start with

$$\text{Premium} = \frac{\text{Discounted Losses} + \text{Discounted Fixed Expenses}}{\text{Premium Delay} - \text{Variable Expense Rate} - \text{Profit}}$$

With this, we have

$$\begin{aligned} \text{Premium Delay} &- \text{Variable Expense Rate} - \text{Profit} \\ &= \frac{\text{Discounted Losses}}{\text{Premium}} + \frac{\text{Discounted Fixed Expenses}}{\text{Premium}} \\ \text{Premium Delay} &- \text{Variable Expense Rate} - \text{Profit} - \frac{\text{Discounted Fixed Expenses}}{\text{Premium}} \\ &= \frac{\text{Discounted Losses}}{\text{Premium}} \\ &= \text{Target Discounted Loss Ratio}. \end{aligned}$$

For simplification, we will not repeat that the components of the rate change factor are discounted. In the loss ratio method, we compare the projected loss ratio to the target loss ratio. A projected loss ratio that exceeds the target loss ratio implies the need for a rate increase. A projected loss ratio that is less than the target loss ratio implies the need for a rate decrease.

### 10.10.3 Experience Period Loss Ratios

Earlier in this section, we described the required adjustments to estimate premiums. We apply those same adjustments to the experience period loss ratios.

### 10.10.4 Adjustments to Loss

- As with the development of pure premiums described above, actuaries will typically review claims experience over a multi-year (typically three to seven years) period to smooth the year-to-year randomness. The years in the experience period are similarly weighted to balance responsiveness to more recent experience and the stability of a longer-term period.

- The numerator of the loss ratio will be *ultimate losses*.
- We will consider the presence of catastrophe and large losses in the claims experience.
- We need to adjust the experience period losses to the cost level of the proposed rate program. We discussed this trend adjustment in Section 10.3.2. We apply the trend factor from the average accident date of the experience period to the average accident date of the proposed rate program. For example, if we are estimating rates that will underlie twelve month policies written in calendar year 2025, the average accident date of the prospective rate program will be 31 December 2025 (sometimes rounded to 1 January 2026). The first policy of the prospective period will be written on 1 January 2025 and expire on 31 December 2025. Assuming even distribution of claim events during the policy, the average accident date (midpoint) of that policy is 1 July 2025. Correspondingly, the last policy of the prospective period will be written on 31 December 2025 and expire on 31 December 2026 with an average accident date (midpoint) of that policy is 1 July 2026. Therefore, the midpoint of all policies written under the proposed rate program is 31 December 2025. To adjust experience for accident year 2022, we apply 3.5 years of trend. The average accident date of accident year 2022 is 1 July 2022 - so 3.5 years is the distance in time to the average accident date of the proposed rate program.

#### 10.10.5 Premium On-Level Adjustment

We also need to adjust premiums for the effect of rate changes. We refer to this adjustment as “on-leveling.” There are two common approaches to on-leveling.

- The Parallelogram Method: Premium on-level factors use historical rate change calculations. For example, if the company adopted a +10% rate change on 1 July 2022, then the 2022 earned premium would need to be adjusted by +7.5%. - Policies written prior to 1 July 2022 would need to be adjusted by +10%; - For the premium earned after 1 July 2022, half would be earned on policies written under the old rate levels and require the 10% adjustment and half would be written on policies written under the higher rate levels and require no adjustment. The weighted average of these adjustments is +7.5%.
- Extension of Exposures: The extension of exposures method is a more detailed approach which involves the re-rating of all historical policies at current rates. It is more precise as the parallelogram method relies on rate changes that were calculated as the average rate change given the mix of business at that time. However, the mix of business may change and the rate

change effect on the current mix may be different. The extension of exposures does not rely on those average rate changes and instead relies only on current rates.

#### 10.10.6 Premium Trend

Experience period premiums must also be adjusted for premium trend, and the basis of premium must match the loss trend. For example, insureds may purchase higher limits of coverage to protect against higher inflation. These higher limits would be reflected in the internal claims experience and may underlie the data used to measure loss trend. If we are considering these changes in the loss trend, then we also need to consider the effect of higher limit purchases in premium trend.

#### 10.10.7 Credibility

Oftentimes, the experience being reviewed is not “fully credible.” That is, the predictive value of the data is limited. We, therefore, need to consider an alternative indication of the projected loss ratio to calculate a credibility-weighted loss ratio. We refer to this alternative indicator as the complement of credibility. A common complement is the net loss trend (loss trend/premium trend). The assumption underlying the use of net loss trend as a complement is that in the absence of an alternative indication, we would need to adjust the rate level to consider changes in cost level. Chapter ?? describes credibility in detail.

**Example. Loss Ratio Indicated Change Factor.** Assume the following information:

- Experience period loss and LAE ratio = 65%
- Experience period credibility = 80%
- Loss Trend = 5%
- Premium On-Level Adjustment = 1.075
- Premium Trend = 2%
- Premium Delay Factor = 0.99
- Projected fixed expense ratio = 6.5%
- Variable expense = 25%
- Target UW profit = 6.1% .

With these assumptions, the indicated change factor can be calculated as

$$\text{Experience Period Loss Ratio} = 65\% \times \frac{1.05}{1.075 \times 1.02} = 62.2\%$$

$$\text{Target Loss Ratio} = 0.99 - 6.5\% - 25\% - 6.1\% = 61.4\%$$

$$\text{Complement of Credibility} = 0.614 \times \frac{1.05}{1.02} = 63.2\%$$

$$\text{Credibility-weighted loss ratio} = 62.2\% \times 80\% + 63.2\% \times (1 - 80\%) = 62.4\%$$

$$\text{Indicated loss ratio} = 62.4\% / 61.4\% = 1.016.$$

This means that overall average rate level should be increased by 1.6%.

---

## 10.11 Further Resources and Contributors

This chapter serves as a bridge between the technical introduction of this book and an introduction to pricing and ratemaking for practicing actuaries. For readers interested in learning practical aspects of pricing, we recommend introductions by the Society of Actuaries in [Friedland \(2013\)](#) and by the Casualty Actuarial Society in [Werner and Modlin \(2016\)](#). For a classic risk management introduction to pricing, see [Niehaus and Harrington \(2003\)](#). See also [Finger \(2006\)](#) and [Frees \(2014\)](#).

[Bühlmann \(1985\)](#) was the first in the academic literature to argue that pricing should be done first at the portfolio level (he referred to this as a *top down* approach) which would be subsequently reconciled with pricing of individual contracts. See also the discussion in [Kaas et al. \(2008\)](#), Chapter 5.

For more background on pricing principles, a classic treatment is by [Gerber \(1979\)](#) with a more modern approach in [Kaas et al. \(2008\)](#). For more discussion of pricing from a financial economics viewpoint, see [Bauer et al. \(2013\)](#).

- **Edward (Jed) Frees**, University of Wisconsin-Madison, and **José Garrido**, Concordia University were the principal authors of the initial version of this chapter.
  - Chapter reviewers included Chun Yong Chew, Curtis Gary Dean, Brian Hartman, and Jeffrey Pai.
- **Rajesh Sahasrabuddhe**, Oliver Wyman, is the author of the second edition of this chapter. Email: [rajesh1004@gmail.com](mailto:rajesh1004@gmail.com) for chapter comments and suggested improvements.

### TS 10.A. Rate Regulation

Insurance regulation helps to ensure the financial stability of insurers and to protect consumers. Insurers receive premiums in return for promises to pay

in the event of a contingent (insured) event. Like other financial institutions such as banks, there is a strong public interest in promoting the continuing viability of insurers.

### Market Conduct

To help protect consumers, regulators impose administrative rules on the behavior of market participants. These rules, known as market conduct regulation, provide systems of regulatory controls that require insurers to demonstrate that they are providing fair and reliable services, including rating, in accordance with the statutes and regulations of a jurisdiction.

1. *Product regulation* serves to protect consumers by ensuring that insurance policy provisions are reasonable and fair, and do not contain major gaps in coverage that might be misunderstood by consumers and leave them unprotected.
2. The insurance product is the insurance contract (policy) and the coverage it provides. Insurance contracts are regulated for these reasons:
  - a. Insurance policies are complex legal documents that are often difficult to interpret and understand.
  - b. Insurers write insurance policies and sell them to the public on a “take it or leave it” basis.

Market conduct includes rules for *intermediaries* such as agents (who sell insurance to individuals) and brokers (who sell insurance to businesses). Market conduct also includes *competition policy regulation*, designed to ensure an efficient and competitive marketplace that offers low prices to consumers.

### Rate Regulation

*Rate regulation* helps guide the development of premiums and so is the focus of this chapter. As with other aspects of market conduct regulation, the intent of these regulations is to ensure that insurers not take unfair advantage of consumers. Rate (and policy form) regulation is common worldwide.

The amount of regulatory scrutiny varies by insurance product. Rate regulation is uncommon in life insurance. Further, in non-life insurance, most commercial lines and reinsurance are free from regulation. Rate regulation is common in automobile insurance, health insurance, workers compensation, medical malpractice, and homeowners insurance. These are markets in which insurance is mandatory or in which universal coverage is thought to be socially desirable.

There are three principles that guide rate regulation: rates should

- be adequate (to maintain insurance company solvency),
- but not excessive (not so high as to lead to exorbitant profits),
- nor unfairly discriminatory (price differences must reflect expected claim and expense differences).

Recently, in auto and home insurance, the twin issues of availability and affordability, which are not explicitly included in the guiding principles, have been assuming greater importance in regulatory decisions.

### Rates are Not Unfairly Discriminatory

Some government regulations of insurance restrict the amount, or level, of premium rates. These are based on the first two of the three guiding rate regulation principles, that rates be adequate but not excessive. This type of regulation is discussed further in the following section on types of rate regulation.

Other government regulations restrict the type of information that can be used in risk classification. These are based on the third guiding principle, that rates not be unfairly discriminatory. “Discrimination” in an insurance context has a different meaning than commonly used; for our purposes, discrimination means the ability to distinguish among things or, in our case, policyholders. The real issue is what is meant by the adjective “fair.”

In life insurance, it has long been held that it is reasonable and fair to charge different premium rates by age. For example, a life insurance premium differs dramatically between an 80 year old and someone aged 20. In contrast, it is unheard of to use rates that differ by:

- ethnicity or race,
- political affiliation, or
- religion.

It is not a matter of whether data can be used to establish statistical significance among the levels of any of these variables. Rather, it is a societal decision as to what constitutes notions of “fairness.”

Different jurisdictions have taken different stances on what constitutes a fair rating variable. For example, in some jurisdictions for some insurance products, gender is no longer a permissible variable. As an illustration, the European Union now prohibits the use of gender for automobile rating. As another example, in the U.S., many discussions have revolved around the use of credit ratings to be used in automobile insurance pricing. Credit ratings are designed to measure consumer financial responsibility. Yet, some argue that credit scores are good proxies for ethnicity and hence should be prohibited.

In an age where more data is being used in imaginative ways, discussions of what constitutes a fair rating variable will only become more important going forward and much of that discussion is beyond the scope of this text. However, it is relevant to the discussion to remark that actuaries and other data analysts can contribute to societal discussions on what constitutes a “fair” rating variable in unique ways by establishing the magnitude of price differences when using variables under discussion.

### Types of Rate Regulation

There are several methods, that vary by the level of scrutiny, by which regulators may restrict the rates that insurers offer.

The most restrictive is a government prescribed regulatory system, where the government regulator determines and promulgates the rates, classifications, forms, and so forth, to which all insurers must adhere. Also restrictive are prior approval systems. Here, the insurer must file rates, rules, and so forth, with government regulators. Depending on the statute, the filing becomes effective when a specified waiting period elapses (if the government regulator does not take specific action on the filing, it is deemed approved automatically) or when the government regulator formally approves the filing.

The least restrictive is a no file or *record maintenance* system where the insurer need not file rates, rules, and so forth, with the government regulator. The regulator may periodically examine the insurer to ensure compliance with the law. Another relatively flexible system is the file only system, also known as *competitive* rating, where the insurer simply keeps files to ensure compliance with the law.

In between these two extremes are the (1) file and use, (2) use and file, (3) modified prior approval, and (4) flex rating systems.

1. File and Use: The insurer must file rates, rules, and so forth, with the government regulator. The filing becomes effective immediately or on a future date specified by the filer.
2. Use and File: The filing becomes effective when used. The insurer must file rates, rules, and so forth, with the government regulator within a specified time period after first use.
3. Modified Prior Approval: This is a hybrid of “prior approval” and “file and use” laws. If the rate revision is based solely on a change in loss experience then “file and use” may apply. However, if the rate revision is based on a change in expense relationships or rate classifications, then “prior approval” may apply.
4. Flex (or Band) Rating: The insurer may increase or decrease a rate

within a “flex band,” or range, without approval of the government regulator. Generally, either “file and use” or “use and file” provisions apply.

---

For a broad introduction to government insurance regulation from a global perspective, see the website of the [International Association of Insurance Supervisors \(IAIS\)](#).



---

## Bibliography

---

- Aalen, Odd (1978). “Nonparametric inference for a family of counting processes,” *The Annals of Statistics*, Vol. 6, pp. 701–726.
- Abbott, Dean (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*, Hoboken, NJ. Wiley.
- Abdullah, Mohammad F. and Kamsuriah Ahmad (2013). “The mapping process of unstructured data to structured data,” in *2013 International Conference on Research and Innovation in Information Systems (ICRIIS)*, pp. 151–155.
- Actuarial Community (2025). *Loss Data Analytics*, URL: <https://openacttets.github.io/Loss-Data-Analytics/index.html>.
- Aggarwal, Charu C. (2015). *Data Mining: The Textbook*, New York, NY. Springer.
- Bailey, Robert A. and J. Simon LeRoy (1960). “Two studies in automobile ratemaking,” *Proceedings of the Casualty Actuarial Society Casualty Actuarial Society*, Vol. XLVII.
- Bauer, Daniel, Richard D. Phillips, and George H. Zanjani (2013). “Financial pricing of insurance,” in *Handbook of Insurance*. Springer, pp. 627–645.
- Bégin, Jean-François (2019). “Economic Scenario Generator and Parameter Uncertainty: A Bayesian Approach,” *ASTIN Bulletin*, Vol. 49, pp. 335–372.
- (2021). “On Complex Economic Scenario Generators: Is Less More?” *ASTIN Bulletin*, Vol. 51, pp. 779–812.
- (2023). “Ensemble Economic Scenario Generators: Unity Makes Strength,” *North American Actuarial Journal*, Vol. 27, pp. 444–471.
- Bermúdez, Lluís and Dimitris Karlis (2011). “Bayesian Multivariate Poisson Models for Insurance Ratemaking,” *Insurance: Mathematics and Economics*, Vol. 48, pp. 226–236.
- Bernardo, José M and Adrian FM Smith (2009). *Bayesian Theory*. John Wiley & Sons: New York, NY, United States of America.
- Bignozzi, Valeria and Andreas Tsanakas (2016). “Parameter Uncertainty and

- Residual Estimation Risk," *Journal of Risk and Insurance*, Vol. 83, pp. 949–978.
- Billingsley, Patrick (2008). *Probability and measure*. John Wiley & Sons.
- Bishop, Christopher M. (2007). *Pattern Recognition and Machine Learning*, New York, NY. Springer.
- Bowers, Newton L., Hans U. Gerber, James C. Hickman, Donald A. Jones, and Cecil J. Nesbitt (1986). *Actuarial Mathematics*. Society of Actuaries Itasca, Ill.
- Box, George E. P. (1980). "Sampling and Bayes' inference in scientific modelling and robustness," *Journal of the Royal Statistical Society. Series A (General)*, pp. 383–430.
- Breiman, Leo (2001). "Statistical Modeling: The Two Cultures," *Statistical Science*, Vol. 16, pp. 199–231.
- Breiman, Leo, Jerome Friedman, Charles J. Stone, and R.A. Olshen (1984). *Classification and Regression Trees*, Raton Boca, FL. Chapman and Hall/CRC.
- Bühlmann, Hans (1985). "Premium calculation from top down," *ASTIN Bulletin: The Journal of the IAA*, Vol. 15, pp. 89–101.
- Buttrey, Samuel E. and Lyn R. Whitaker (2017). *A Data Scientist's Guide to Acquiring, Cleaning, and Managing Data in R*, Hoboken, NJ. Wiley.
- Cairns, Andrew JG (2000). "A Discussion of Parameter and Model Uncertainty in Insurance," *Insurance: Mathematics and Economics*, Vol. 27, pp. 313–330.
- Cairns, Andrew JG, David Blake, and Kevin Dowd (2006). "A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration," *Journal of Risk and Insurance*, Vol. 73, pp. 687–718.
- Chen, Min, Shiwen Mao, Yin Zhang, and Victor CM Leung (2014). *Big Data: Related Technologies, Challenges and Future Prospects*, New York, NY. Springer.
- Cheung, Eric CK, Weihong Ni, Rosy Oh, and Jae-Kyung Woo (2021). "Bayesian Credibility Under a Bivariate Prior on the Frequency and the Severity of Claims," *Insurance: Mathematics and Economics*, Vol. 100, pp. 274–295.
- Cowles, Mary Kathryn (2013). *Applied Bayesian Statistics: With R and Open-*

- BUGS Examples.* Springer Science & Business Media: New York, NY, United States of America.
- Cummins, J. David and Richard A. Derrig (2012). *Managing the Insolvency Risk of Insurance Companies: Proceedings of the Second International Conference on Insurance Solvency*, Vol. 12. Springer Science & Business Media.
- Daroczi, Gergely (2015). *Mastering Data Analysis with R*, Birmingham, UK. Packt Publishing.
- De Jong, Piet and Gillian Z. Heller (2008). *Generalized linear models for insurance data*. Cambridge University Press, Cambridge.
- Derrig, Richard A, Krzysztof M Ostaszewski, and Grzegorz A Rempala (2001). “Applications of resampling methods in actuarial practice,” in *Proceedings of the Casualty Actuarial Society*, Vol. 87, pp. 322–364, Casualty Actuarial Society.
- Dickson, David C. M., Mary Hardy, and Howard R. Waters (2013). *Actuarial Mathematics for Life Contingent Risks*. Cambridge University Press.
- Earnix (2013). “2013 Insurance Predictive Modeling Survey,” Earnix and Insurance Services Office, Inc. URL: <https://www.verisk.com/archived/2013/majority-of-north-american-insurance-companies-use-predictive-analytics-to-enhance-business-performance-new-earnix-iso-survey-shows/>, [Retrieved on July 23, 2020].
- Efron, Bradley (1979). “Bootstrap methods: Another look at the bootstrap,” *The Annals of Statistics*, Vol. 7, pp. 1–26.
- (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM.
- (1992). *Bootstrap Methods: Another Look at the Jackknife*, pp. 569–593. Springer New York, URL: [https://doi.org/10.1007/978-1-4612-4380-9\\_41](https://doi.org/10.1007/978-1-4612-4380-9_41), DOI: [http://dx.doi.org/10.1007/978-1-4612-4380-9\\_41](http://dx.doi.org/10.1007/978-1-4612-4380-9_41).
- Fellingham, Gilbert W, Athanasios Kottas, and Brian M Hartman (2015). “Bayesian Nonparametric Predictive Modeling of Group Health Claims,” *Insurance: Mathematics and Economics*, Vol. 60, pp. 1–10.
- Finger, Robert J. (2006). “Risk classification,” pp. 231–276.
- Forte, Rui Miguel (2015). *Mastering Predictive Analytics with R*, Birmingham, UK. Packt Publishing.
- Frees, Edward W (2009). *Regression Modeling with Actuarial and Financial*

- Applications*. Cambridge University Press, URL: <https://doi.org/10.1017/CBO9780511814372>.
- (2014). “Frequency and severity models,” in Edward W Frees, Glenn Meyers, and Richard Derrig eds. *Predictive Modeling Applications in Actuarial Science*, Vol. 1, pp. 138–164. Cambridge University Press Cambridge, URL: <https://doi.org/10.1017/CBO9781139342674>.
- (2015). “Analytics of insurance markets,” *Annual Review of Financial Economics*, Vol. 7, pp. 253–277.
- Frees, Edward W and Lisa Gao (2019). “Predictive Analytics and Medical Malpractice,” *North American Actuarial Journal*, pp. 1–17, URL: <https://doi.org/10.1080/10920277.2019.1634597>, DOI: <http://dx.doi.org/10.1080/10920277.2019.1634597>.
- Frees, Edward W and Fei Huang (2021). “The discriminating (pricing) actuary,” *North American Actuarial Journal*, pp. 1–23, URL: <https://www.tandfonline.com/doi/pdf/10.1080/10920277.2021.1951296>.
- Frees, Edward W, Gee Lee, and Lu Yang (2016). “Multivariate frequency-severity regression models under insurance,” *Risks*, Vol. 4(1), p. 4.
- Frees, Edward W and Emiliano A. Valdez (2008). “Hierarchical insurance claims modeling,” *Journal of the American Statistical Association*, Vol. 103, pp. 1457–1469.
- Friedland, Jacqueline (2013). *Fundamentals of General Insurance Actuarial Analysis*. Society of Actuaries.
- Gan, Guojun (2011). *Data Clustering in C++: An Object-Oriented Approach*, Data Mining and Knowledge Discovery Series, Boca Raton, FL, USA. Chapman & Hall/CRC Press, DOI: <http://dx.doi.org/10.1201/b10814>.
- Gan, Guojun, Chaoqun Ma, and Jianhong Wu (2007). *Data Clustering: Theory, Algorithms, and Applications*, Philadelphia, PA. SIAM Press, DOI: <http://dx.doi.org/10.1137/1.9780898718348>.
- Gelfand, Alan E and Adrian FM Smith (1990). “Sampling-based approaches to calculating marginal densities,” *Journal of the American Statistical Association*, Vol. 85, pp. 398–409.
- Gelman, Andrew and Donald B Rubin (1992). “Inference from iterative simulation using multiple sequences,” *Statistical Science*, pp. 457–472.
- Gerber, Hans U (1979). *An Introduction to Mathematical Risk Theory*, vol. 8 of SS Huebner Foundation Monograph Series. University of Pennsylvania Wharton School SS Huebner Foundation for Insurance Education.

- Goldberger, Arthur S. (1972). "Structural equation methods in the social sciences," *Econometrica: Journal of the Econometric Society*, pp. 979–1001.
- Good, I. J. (1983). "The Philosophy of Exploratory Data Analysis," *Philosophy of Science*, Vol. 50, pp. 283–295.
- Gorman, Mark and Stephen Swenson (2013). "Building believers: How to expand the use of predictive analytics in claims," SAS, URL: <https://www.the-digital-insurer.com/wp-content/uploads/2014/10/265-wp-59831.pdf>, [Retrieved on July 23, 2020].
- Greenwood, Major (1926). "The errors of sampling of the survivorship tables," in *Reports on Public Health and Statistical Subjects*, Vol. 33. London: Her Majesty's Stationery Office.
- Hartman, Brian M and Chris Groendyke (2013). "Model Selection and Averaging in Financial Risk Management," *North American Actuarial Journal*, Vol. 17, pp. 216–228.
- Hartman, Brian M and Matthew J Heaton (2011). "Accounting for Regime and Parameter Uncertainty in Regime-Switching Models," *Insurance: Mathematics and Economics*, Vol. 49, pp. 429–437.
- Hashem, Ibrahim Abaker Targio, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan (2015). "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, Vol. 47, pp. 98 – 115.
- Hastie, Trevor, Robert Tibshirani, and Jerome H Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2. Springer.
- Hastings, WK (1970). "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, Vol. 57, pp. 97–109.
- Haueter, Niels Viggo (2017). "A History of UK Insurance," Technical report.
- Heckman, Philip E and Glenn G Meyers (1983). "The Calculation of Aggregate Loss Distributions from Claim Severity and Claim Count Distributions," in *Proceedings of the Casualty Actuarial Society*, Vol. 70, pp. 49–66.
- Hoerl, Arthur E and Robert W. Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, Vol. 12, pp. 55–67.
- Hogg, Robert V, Elliot A Tanis, and Dale L Zimmerman (2015). *Probability and Statistical Inference, 9th Edition*. Pearson, New York.

- Hox, Joop J. and Hennie R. Boeije (2005). “Data collection, primary versus secondary,” in *Encyclopedia of social measurement*. Elsevier, pp. 593 – 599.
- Huang, Yifan and Shengwang Meng (2020). “A Bayesian Nonparametric Model and its Application in Insurance Loss Prediction,” *Insurance: Mathematics and Economics*, Vol. 93, pp. 84–94.
- Inmon, W.H. and Dan Linstedt (2014). *Data Architecture: A Primer for the Data Scientist: Big Data, Data Warehouse and Data Vault*, Cambridge, MA. Morgan Kaufmann.
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An introduction to statistical learning*, Vol. 112. Springer.
- Janert, Philipp K. (2010). *Data Analysis with Open Source Tools*, Sebastopol, CA. O'Reilly Media.
- Kaas, Rob, Marc Goovaerts, Jan Dhaene, and Michel Denuit (2008). *Modern actuarial risk theory: using R*, Vol. 128. Springer Science & Business Media.
- Kaplan, Edward L. and Paul Meier (1958). “Nonparametric estimation from incomplete observations,” *Journal of the American statistical association*, Vol. 53, pp. 457–481.
- Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot (2012). *Loss Models: From Data to Decisions*. John Wiley & Sons.
- Kreer, Markus, Ayşe Kızılersü, Anthony W Thomas, and Alfredo D Egídio dos Reis (2015). “Goodness-of-fit tests and applications for left-truncated Weibull distributions to non-life insurance,” *European Actuarial Journal*, Vol. 5, pp. 139–163.
- Levin, Bruce, James Reeds et al. (1977). “Compound multinomial likelihood functions are unimodal: Proof of a conjecture of IJ Good,” *The Annals of Statistics*, Vol. 5, pp. 79–87.
- McDonald, James B (1984). “Some generalized functions for the size distribution of income,” *Econometrica: journal of the Econometric Society*, pp. 647–663.
- McDonald, James B and Yexiao J Xu (1995). “A generalization of the beta distribution with applications,” *Journal of Econometrics*, Vol. 66, pp. 133–152.
- Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller (1953). “Equation of state calculations by fast computing machines,” *Journal of Chemical Physics*, Vol. 21, pp. 1087–1092.

- Meyers, Glenn (1994). “Quantifying the Uncertainty in Claim Severity Estimates for an Excess Layer When Using the Single Parameter Pareto,” in *Proceedings of the Casualty Actuarial Society*, Vol. 81, pp. 91–122.
- Meyers, Glenn and Nathaniel Schenker (1983). “Parameter Uncertainty in the Collective Risk Model,” *PCAS LXX*, Vol. 111, p. 15.
- Miles, Matthew, Michael Hberman, and Johnny Sdana (2014). *Qualitative Data Analysis: A Methods Sourcebook*, Thousand Oaks, CA. Sage, 3rd edition.
- Mirkin, Boris (2011). *Core Concepts in Data Analysis: Summarization, Correlation and Visualization*, London, UK. Springer.
- Mitchell, Tom M. (1997). *Machine Learning*. McGraw-Hill.
- NAIC Glossary (2018). “Glossary of Insurance Terms,” National Association of Insurance Commissioners, URL: [https://www.naic.org/consumer\\_glossary.htm](https://www.naic.org/consumer_glossary.htm), [Retrieved on Sept 11, 2018].
- Niehaus, Gregory and Scott Harrington (2003). *Risk Management and Insurance*, New York. McGraw Hill.
- O’Donnell, Terence (1936). *History of Life Insurance in its Formative Years*. American Conservation Company: Chicago, IL, United States of America.
- O’Leary, D. E. (2013). “Artificial Intelligence and Big Data,” *IEEE Intelligent Systems*, Vol. 28, pp. 96–99.
- Olkin, Ingram, A John Petkau, and James V Zidek (1981). “A comparison of n estimators for the binomial distribution,” *Journal of the American Statistical Association*, Vol. 76, pp. 637–642.
- Picard, Richard R. and Kenneth N. Berk (1990). “Data splitting,” *The American Statistician*, Vol. 44, pp. 140–147.
- Pries, Kim H. and Robert Dunnigan (2015). *Big Data Analytics: A Practical Guide for Managers*, Boca Raton, FL. CRC Press.
- Quenouille, Maurice H (1949). “Approximate tests of correlation in time-series,” *Journal of the Royal Statistical Society. Series B*, Vol. 11, pp. 68–84.
- Robert, Christian P and George Casella (1999). *Monte Carlo Statistical Methods*. Springer: New York, NY, United States of America.
- Ruppert, David, Matt P Wand, and Raymond J Carroll (2003). *Semiparametric regression*, No. 12. Cambridge University Press.

- Samuel, A. L. (1959). "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, Vol. 3, pp. 210–229.
- Shmueli, Galit (2010). "To Explain or to Predict?" *Statistical Science*, Vol. 25, pp. 289–310.
- Snee, Ronald D. (1977). "Validation of regression models: methods and examples," *Technometrics*, Vol. 19, pp. 415–428.
- Stigler, Stephen M (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.
- Tevet, Dan (2016). "Applying Generalized Linear Models to Insurance Data," *Predictive Modeling Applications in Actuarial Science: Volume 2, Case Studies in Insurance*, p. 39.
- The Organization for Economic Cooperation and Development (OECD) (2021). "OECD Insurance Statistics 2021," OECD iLibrary, URL: [https://read.oecd-ilibrary.org/finance-and-investment/oecd-insurance-statistics-2021\\_841fa619-en#page1](https://read.oecd-ilibrary.org/finance-and-investment/oecd-insurance-statistics-2021_841fa619-en#page1), [Retrieved on 1 August, 2022].
- Tukey, John W. (1962). "The Future of Data Analysis," *The Annals of Mathematical Statistics*, Vol. 33, pp. 1–67.
- de Valpine, Perry, Daniel Turek, Christopher J Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang, and Rastislav Bodik (2017). "Programming with models: Writing statistical algorithms for general model structures with nimble," *Journal of Computational and Graphical Statistics*, Vol. 26, pp. 403–413.
- Vats, Dootika, James M Flegal, and Galin L Jones (2019). "Multivariate output analysis for Markov chain Monte Carlo," *Biometrika*, Vol. 106, pp. 321–337.
- Venter, Gary (1983). "Transformed beta and gamma distributions and aggregate losses," in *Proceedings of the Casualty Actuarial Society*, Vol. 70, pp. 289–308.
- Werner, Geoff and Claudine Modlin (2016). *Basic Ratemaking, Fifth Edition*. Casualty Actuarial Society, URL: [https://www.casact.org/library/studynotes/werner\\_modlin\\_ratemaking.pdf](https://www.casact.org/library/studynotes/werner_modlin_ratemaking.pdf), [Retrieved on April 1, 2019].
- Wolny-Dominiak, Alicja and Michal Trzesiok (2014). "Package ‘insuranceData’," Technical report, The Comprehensive R Archive Network.
- Young, Virginia R (2014). "Premium principles," *Wiley StatsRef: Statistics Reference Online*.